This CVPR2015 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild

Xiangyu Zhu Zhen Lei* Junjie Yan Dong Yi Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun Donglu, Beijing 100190, China.
{xiangyu.zhu,zlei,jjyan,dony.yi,szli}@nlpr.ia.ac.cn

Abstract

Pose and expression normalization is a crucial step to recover the canonical view of faces under arbitrary conditions, so as to improve the face recognition performance. An ideal normalization method is desired to be automatic, database independent and high-fidelity, where the face appearance should be preserved with little artifact and information loss. However, most normalization methods fail to satisfy one or more of the goals. In this paper, we propose a High-fidelity Pose and Expression Normalization (HPEN) method with 3D Morphable Model (3DMM) which can automatically generate a natural face image in frontal pose and neutral expression. Specifically, we firstly make a landmark marching assumption to describe the non-correspondence between 2D and 3D landmarks caused by pose variations and propose a pose adaptive 3DMM fitting algorithm. Secondly, we mesh the whole image into a 3D object and eliminate the pose and expression variations using an identity preserving 3D transformation. Finally, we propose an inpainting method based on Possion Editing to fill the invisible region caused by self occlusion. Extensive experiments on Multi-PIE and LFW demonstrate that the proposed method significantly improves face recognition performance and outperforms state-of-the-art methods in both constrained and unconstrained environments.

1. Introduction

During the past decade, face recognition has attracted much attention due to its great potential value in real world applications, such as access control, identity verification and video surveillance. However, in unconstrained environment the performance of face recognition always drops significantly because of large variations caused by pose, illumination, expression, occlusion and so on. Among them pose and expression have always been important challenges because they can dramatically increase intra-person variances, sometimes even exceeding inter-person variances. To deal with the two challenges, many promising works have been developed, which can be divided into two categories: feature level normalization and image level normalization.

The feature level normalization aims at designing face representations with robustness to pose and expression variations [21, 32, 44, 35, 42]. For instance, the Pose Adaptive Filter [48] adjusts its filter according to pose conditions and extracts features on semantic consistent positions. The High-dim LBP [21] concatenates many local descriptors to a high-dim form and demonstrates robustness to global and local distortions. Besides hand crafted features, discriminative features can also be learned from data. Fisher vector [41], Learning Based Descriptor [17] and Probabilistic Elastic Matching [32] use unsupervised learning techniques to learn encoders from training examples. Convolutional Neural Network (CNN) provides a framework to learn face representations in a supervised form, and has achieved significant improvements in recent years [44, 46].

The image level normalization aims to synthesize a virtual canonical-view and expression-free image from one under arbitrary conditions. The advantage of this category is that it can be easily incorporated into traditional face recognition framework as a pre-processing procedure. There are 2D and 3D methods. One type of 2D methods estimates a spatial mapping (a flow), either pixel-wise or patch-wise, to simulate the geometry transformation in 3D space, such as Stack Flow [4], Markov Random Field [3] and Morphable Displacement [34]. In these methods, although the face pixels are rearranged to the frontal view, the shape and consistency are not well preserved. Another type of 2D methods tries to learn the appearance transformations between different poses, such as Local Linear Regression [18] and Tied Factor [39]. These methods use linear models to approxi-

*Corresponding author.

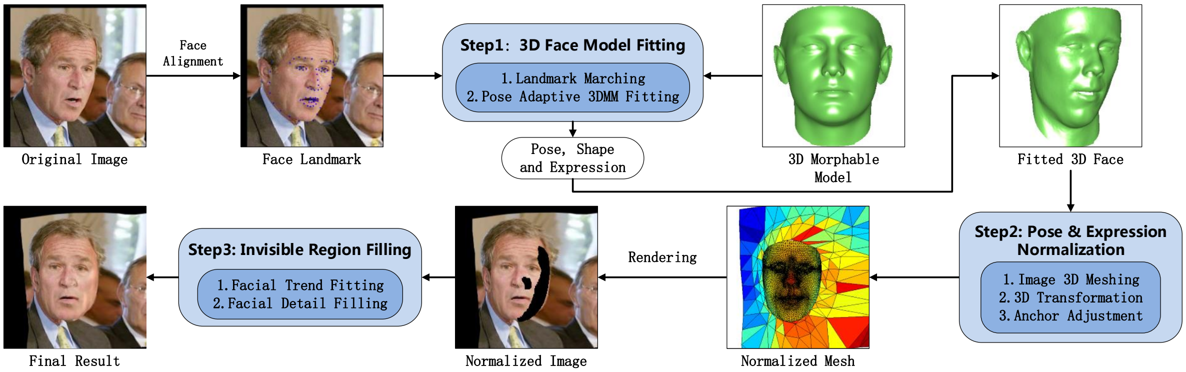

Figure 1. Overview of the High-Fidelity Pose and Expression Normalization (HPEN) method

mate the highly non-linear pose variations and thus cannot always retain the identity information. Recently, FIP [50] and SPAE [29] train high-dimensional non-linear appearance transformations with deep models, achieving state-of-the-art performance. However, it is necessary to prepare a very large and carefully designed database, where a gallery image under neutral condition is provided for each subject. This requirement is not always satisfied in real application.

Since pose variations are caused by 3D rigid transformations of face, 3D methods are inherently more intuitive and accurate. 3D methods estimate the depth information with a 3D model and normalize faces through 3D transformations. A representative method is the 3D Morphable Model (3DMM) [10], which constructs its 3D shape and texture model with PCA and estimates model parameters by minimizing the difference between image and model appearance. Although proposed a decade before, 3DMM still has competitive performance [29]. However, this method suffers from the amazing one-minute-per-image time cost. An alternative method is the landmark based 3D face model fitting [5, 1, 46, 13, 12, 38], which estimates the model parameters with the correspondence between 2D and 3D landmarks. This method is very efficient but suffers from the problem that the semantic positions of face contour landmarks differ from pose to pose. Besides, most 3D normalization methods do not fill the invisible region caused by self-occlusion, leading to large artifacts and non face-like normalization results [5, 23].

In this paper, we present a pose and expression normalization method to recover the canonical-view, expression-free image with "high fidelity", which indicates preserving the face appearance with little artifact and information loss. The contributions are as follows: Firstly, we make a "landmark marching" assumption to describe the movement of 3D landmarks across poses and propose a landmark based pose adaptive 3DMM fitting method (Section 2). Secondly, we propose a identity preserving normalization by meshing the whole image into a 3D object and normalizing it with 3D transformations (Section 3). Finally, we propose a "Trend Fitting and Detail Filling" method to fill the invisible region with poisson editing, leading to smooth and natural normalization result. Based on the well developed landmark detector [47], the entire normalization system does not contain any learning procedure, leading to good generalization performance to different environments. The proposed method is briefly summarized in Fig. 1. The code can be downloaded from `http://www.cbsr.ia.ac.cn/users/xiangyuzhu`.

2. Pose Adaptive 3DMM Fitting

In this section, we firstly introduce the 3D Morphable Model (3DMM) and then describe our pose adaptive 3DMM fitting method.

2.1. 3D Morphable Model

3D Morphable Model is one of the most successful methods to describe the 3D face space. Constructed by linear combinations of face scans, 3DMM can approximate arbitrary face shape to a considerable extent. Recently, Chu et al. [23] extend 3DMM to contain expressions as the offset to the neutral face.

$$S = \overline{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \quad (1)$$

where S is the 3D face, $\overline{S}$ is the mean shape, A_{id} is the principle axes trained on the 3D face scans with neutral expression and α_{id} is the shape weight, A_{exp} is the principle axes trained on the offset between expression scans and neutral scans and α_{exp} represents the expression weight. In this work, we merge two popular face models with Nonrigid ICP [2] to construct our 3DMM. The identity shape A_{id} comes from the Basel Face Model (BFM) [36] and the expression A_{exp} comes from the Face Warehouse [14].

To fit 3DMM to a face image, we project the face model onto the image plane with the Weak Perspective Projection:

$$s_{2d} = fPR(\alpha, \beta, \gamma)(S + t_{3d}) \quad (2)$$

where s_{2d} is the 2D positions of 3D points on the image plane, f is the scale factor, P is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $R(\alpha, \beta, \gamma)$ is the 3×3 rotation matrix constructed with pitch(α), yaw(β) and roll(γ) and t_{3d} is the translation vector. The fitting process needs to search the ground truth 2D coordinates s_{2dt} of 3D points and estimate the model parameters by minimizing the distance between s_{2d} and s_{2dt}:

$$\arg \min_{f, R, t_{3d}, \alpha_{id}, \alpha_{exp}} \|s_{2dt} - s_{2d}\| \tag{3}$$

2.2. Landmark Marching

Benefit from the current breakthrough of face alignment algorithm [47, 16], robustly detecting face landmarks in unconstrained environment has become possible. If we mark the corresponding 3D landmarks on the face model, a sparse correspondence between 3D and 2D space can be constructed. Then 3DMM can be fitted with Eqn. (3), where the s_{2dt} and s_{2d} are the 2D and projected 3D landmarks respectively. However, this fitting framework has a big problem that the landmarks on the cheek boundary are not consistent across poses. When faces deviate from the frontal pose, the landmarks on the contour will "move" to the face silhouette and break the correspondence, see Fig. 2 for example.

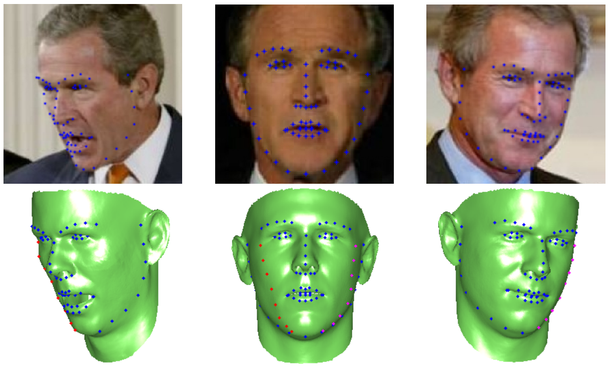

Figure 2. The landmark marching phenomenon. The blue points on the 3D face are standard landmark positions. The red and magenta points are moved landmarks, which are also plotted on the frontal face.

To solve the problem, Lee et al. [31] and Qu et al. [40] detect and discard moved landmarks. This method cannot make full use of landmark constrains. Asthana et al. [6] build a look up table containing 3D landmark configurations for each pose. This method depends on pose estimation and needs a large table in unconstrained environment. In this paper, we intend to localize the moving contour landmarks and rebuild the correspondence automatically. We make an assumption called "landmark marching" to describe the phenomenon: **When pose changes, if a contour landmark is visible, it will not move; or it will move along the parallel to the visibility boundary**. The parallels are shown in Fig. 3(a). In the assumption, we restrict the landmark paths to the parallels and give clear definition of their positions. Note that in the fitting process, pose and landmark configuration depends on each other and should be estimated in an iterative manner. To improve efficiency, we propose an approximation method to avoid iterative visibility estimation. Observing that human head is roughly a cylinder [43] and for a cylinder in any out-of-plane rotation (yaw and pitch), the visibility boundary always corresponds to the generatrix with extreme x coordinates (minimum in left and maximum in right), see Fig. 3(b). Thus in landmark marching, if a parallel crosses the visibility boundary, the point with extreme x will be the marching destination. Inspired by this observation, we first project the 3D face with only yaw and pitch to eliminate in-plane rotation:

$$S_{\alpha,\beta} = R(\alpha, \beta, 0)S \tag{4}$$

Then, for each parallel, the point with extreme x coordinate will be chosen as the adjusted contour landmark, see Fig. 3(c)3(d).

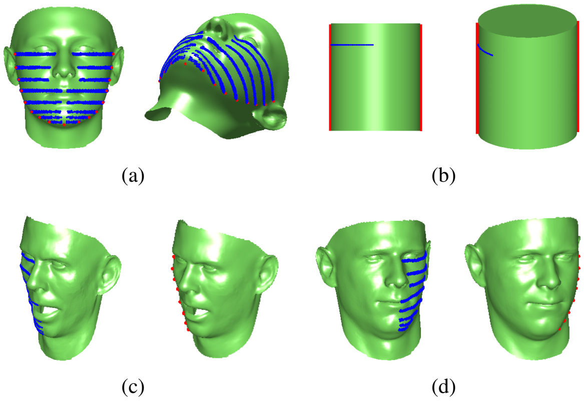

Figure 3. (a) The parallels on the mean face, the red points are the standard landmark positions. (b) The landmark marching on a cylinder. The left one is the frontal view, the right one is rotated with yaw and pitch. The red lines are the generatrix corresponding to the visibility boundary, the blue line is a parallel. (c)(d) Project 3D face with only yaw and pitch and get adjusted landmark positions. The 3D shape of (c) and (d) come from the first and third column of Fig. 2 respectively.

With the landmark marching, the correspondence between 2D and 3D landmarks is rebuilt, and the 3DMM fitting can be summarized as solving the equation:

$$s_{2d\_land} = fPR[\overline{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} + t_{3d}]_{land} \tag{5}$$

where s_{2d_land} is the 2D landmarks and the subscript $land$ means only the adjusted 3D landmarks are selected. Parameters needed to be solved are the shape α_{id}, expression α_{exp}, pose f, R, t_{3d} and landmark configuration $land$. Each group of parameters can be solved when the other three are fixed. In detail, firstly the α_{id} and α_{exp} are initialized to zero and pose is coarsely estimated with only facial feature landmarks using Weak Perspective Projection [11], then landmark marching is conducted to update landmark configuration $land$. After initialization, the parameters are estimated by solving Eqn. (5) in an iterative manner (4 times in this work). Since all steps are linear problems and are only related with landmarks, the fitting is very efficient, which can always finish in less than 0.2s.

3. Identity Preserving Normalization

In this section, we demonstrate how to normalize pose and expression while preserving the identity information. As we know, the shape and texture of a face contain the main identity information and should be kept constant when normalizing. With a fitted 3DMM, we can directly mapping pixels as the face texture and retain the shape parameters during normalization. Besides, the appearance surrounding the face region also contain discriminative information for face recognition [30, 19]. However, most previous works either only keep the internal face region and dropping the information around the face [5, 23, 50, 29] or warp the pixels of surrounding region to the fixed positions so that some shape information is lost [9]. In this work, we propose to estimate the depth information of the whole image and thus the pose and expression can be easily corrected by 3D transformation to preserve as much identity information as possible.

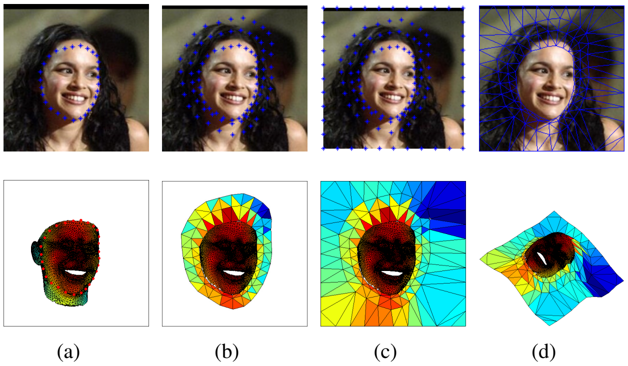

Figure 4. 2D and 3D view of 3D-meshing. (a) The boundary anchors. (b) The surrounding anchors. (c) The background anchors. (d) Triangulation and better view of depth information.

3.1. 3D Meshing and Normalization

In order to ensure the smooth transition from the face region to its background after pose normalization, except face region, we also estimate the depth of the external face region and the background. Specifically, we estimate the depth of anchors from three groups (shown in Fig. 4). One is the boundary anchors which are located on the face contour and adjusted by landmark marching (Fig. 4(a)). The second group is the surrounding anchors which enclose a larger region containing headback, ear and neck (Fig. 4(b)). The depth of these anchors can be approximately estimated by enlarging the 3D face with increasing the scale parameter f and translating the nosetip to the original position. The third is the background anchors located on the image boundary (Fig. 4(c)), and their depth is set to the same as the closest surrounding anchor. Once all anchors are determined, we apply the delaunay algorithm to triangulate anchors and obtain the 3D meshed face object, shown in Fig. 4(d).

After 3D-meshing, the pose can be corrected with the inverse rotation matrix R^{-1}.

$$S_{img\_rn} = R^{-1} S_{img} \tag{6}$$

where S_{img} is the meshed face object containing 3D face model and anchors, see Fig. 4(c), R is the estimated rotation matrix in 3DMM fitting and S_{img_rn} is the rigidly normalized mesh, see Fig. 5(a). For expression normalization, we set the α_{exp} to the neutral expression weight α_{exp_neu} [23], see Fig. 5(b). Note that the shape parameters are kept unchanged to preserve identity information.

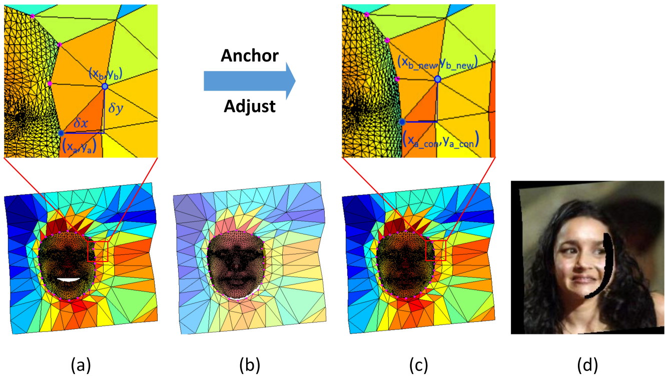

Figure 5. (a) The rigidly normalized mesh, the magenta points are the boundary anchors. (b) The result of pose and expression normalization. There generates a hollow region below the chin due to expression change. We also make mesh transparent to demonstrate the face region occluded by background mesh. (c) The result of anchor adjustment. The boundary anchors move to the predefined face contour positions and all anchors are adjusted. (d) The normalized image where the black region is the invisible region.

3.2. Anchor Adjustment

From Fig. 5(b), one can see that after pose and expression normalization, the semantic of boundary anchors does not correspond to the face contour due to landmark marching and expression change, needing to be further adjusted. Since anchors are related, all the anchors need to be adjusted to preserve the image structure.

We propose a graph to represent the relationships of the anchors with the rigidly normalized mesh such as Fig. 5(a), in which the vertices are the anchor points and the edges are the lines connecting the vertices in the mesh. Each edge represents an anchor-to-anchor relationship:

$$x_a - x_b = \delta x \qquad y_a - y_b = \delta y \tag{7}$$

where (x_a, y_a) and (x_b, y_b) are two connecting anchors, δx and δy are the offsets in x and y coordinates. In anchor adjustment, we move the boundary anchors to the pre-defined positions on 3D face model and try to keep the spatial distance $(\delta x, \delta y)$ unchanged:

$$x_{a\_con} - x_{b\_new} = x_a - x_b \quad y_{a\_con} - y_{b\_new} = y_a - y_b \tag{8}$$

where (x_{a_con}, y_{a_con}) is the predefined face contour position corresponding to a boundary anchor a, (x_{b_new}, y_{b_new}) is the new position of a connecting anchor b, which needs to be solved, (x_a, y_a) and (x_b, y_b) are the coordinates before adjustment. We can adaptively obtain the adjusted positions of the other two groups of anchors by solving the equations for each connecting a and b, which forming an equation list. The solution can be obtained by least squares, see Fig. 5(c). Afterwards, the normalized image can be rendered by the correspondence between the source image and the normalized image provided by the triangles, see Fig. 5(d).

4. Invisible Region Filling

If the yaw angle of face is too large, there may be some regions become invisible due to self-occlusion. Bad filling of the occluded region will lead to large artifacts after normalization and deteriorate recognition performance. In recent works, Asthana et.al [5] and Chu et.al [23] leave the occluded region unfilled, Ding et.al [24] and Li et.al [34] inpaint the region with the mirrored pixels and the gallery face respectively. They cannot generate a coherent face image just like taken under frontal view. Generally, the basic idea of dealing with self-occlusion is utilizing the facial symmetry. However, due to the existence of illumination, facial symmetry cannot always hold. Directly copying pixels will lead to non-smoothness and weird illumination [24]. In this paper, we propose a new way to deal with the invisibility: **Fitting the trend and filling the detail**, which deals with illumination and texture components separately.

4.1. Facial Trend Fitting

We define the facial trend as the illuminated mean face texture, which represents the large scale face appearance. It can be estimated in an efficient manner. For a 3D lambertian object under arbitrary illumination, its appearance can be approximated by the linear combinations of spherical harmonic reflectance bases [49]. These bases are constructed from the surface normal **n** and albedo λ, which are determined by the 3DMM and the mean face texture (Fig. 6(a)) respectively in this paper. By minimizing the difference between the original image and the spherical harmonic reflectance, we can get the illumination parameters:

$$\gamma^* = \arg\min_{\gamma} \|I - B\gamma\| \tag{9}$$

where I is the image pixels corresponding to 3D points as in Fig. 6(b), B is the spherical harmonic reflectance bases and γ is a 9-dimensional illumination vector. Then the facial trend can be represented as $B\gamma^*$, see Fig. 6(c).

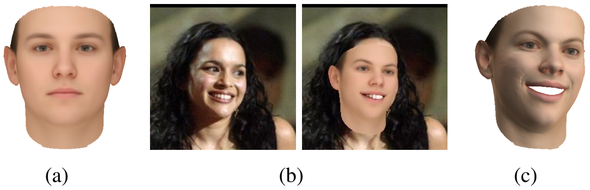

Figure 6. (a) The mean texture. (b) The source image and the projected 3D face. (c) The facial trend.

4.2. Facial Detail Filling

The difference between the image pixel and the facial trend can be seen as the illumination-free facial detail, which roughly satisfies the symmetry assumption and can be estimated with mirroring. In order to further keep the smoothness of filling boundary, we copy gradients instead of pixels. Perez et al. [37] propose an image editing method based on the poisson equation, which can insert a source object into an image seamlessly. The key of poisson image editing is the poisson partial differential equation with Dirichlet boundary condition:

$$\Delta f = w \ \ over \ \ \Omega, \qquad s.t \ \ f|_{\partial\Omega} = f_0|_{\partial\Omega} \tag{10}$$

where f is the edited image to be solved, Δ is the Laplacian operator, w is the Laplacian value of the inserting object, Ω is the editing region, $\partial\Omega$ is the boundary of the region and f_0 is the original image. Setting f_0 as the original detail, w as the laplacian of the mirrored detail, and Ω as the invisible region, the poisson editing can automatically fill the invisible region with great consistency which is guaranteed by the Dirichlet boundary condition. In the end, the facial trend and facial detail are added to form the final result. Fig. 7 demonstrates the process of facial detail filling.

5. Experiments

We evaluate the effectiveness of proposed normalization method in the case of unconstrained (LFW) and constrained (Multi-PIE) face recognition problems, compared with state-of-the-art methods. More normalization results can be found in supplemental material.

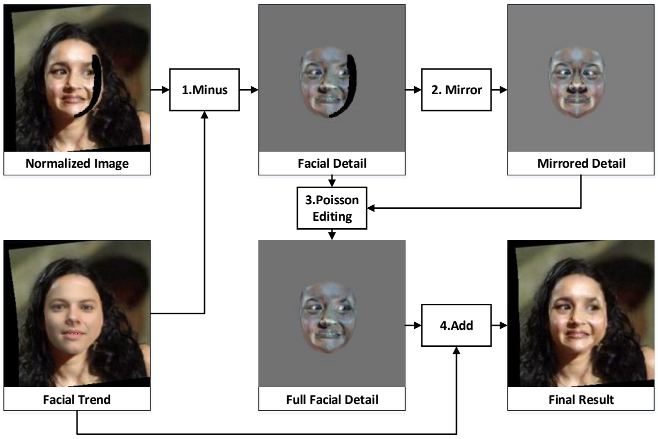

Figure 7. The flowchart of facial detail filling. Step 1: the difference between the face appearance and facial trend is calculated as the detail. Step 2: The facial detail is mirrored to form the inserting object. Step 3: The mirrored detail is inserted into the invisible region (marked with black) with poisson image editing, generating the full facial detail. Step 4: The facial trend and facial detail are added to generate the final result.

5.1. LFW

Labeled Faces in the Wild (LFW) [28] is the most commonly used database for unconstrained face recognition. There are 13223 images from 5729 subjects with variations of expression, pose, occlusion etc. Both "Image restricted" and "Image unrestricted" protocols are adopted in this part. The face recognition performance is reported as the mean accuracy on "View 2" ten splits. During evaluation, we only use the outside data to construct 3DMM (BFM [36] and FaceWarehouse [14]) and train the landmark detector (LFPW [8]). For face recognition, we only use the LFW samples strictly and no outside data is used.

Given a face image, we firstly locate the facial landmarks automatically [47]. The proposed HPEN is then applied to eliminate the pose and expression variations. In HPEN, we conduct invisible region filling only on samples with yaw larger than 10° and directly mirror faces with yaw larger than 40°. For face recognition, we employ the over complete high-dim features [22] including high-dim Gabor (HD-Gabor) and high-dim LBP (HD-LBP) as face representation. The discriminative deep metric learning (DDML) [26] and the joint Bayesian (JB) [20] are used for restricted and unrestricted settings, respectively. The entire normalization process takes about 0.8 seconds on a 3.40GHZ CPU with matlab code.

5.1.1 Performance Analysis

Most of 3D normalization methods [5, 34, 23] only keep the face region and cannot well deal with the invisibility. In this section, we evaluate the benefits from the invisible region and the background. We conduct recognition experiments on three types of images with different level of completeness. The first type "Visible" only keeps visible face region (Fig. 8(b)), the second type "Face" contains complete face region with invisible region filling (Fig. 8(c)) and the last type "Full" is the fully normalized image (Fig. 8(d)).

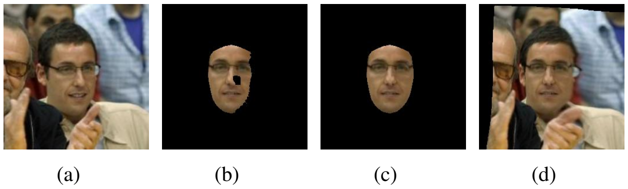

Figure 8. (a) The original image, (b) "Visible", (c) "Face", (d) "Full".

Table 1. The recognition accuracy on three types of images in LFW with different protocols and features.

Setting	Features	Visible	Face	Full
Restricted	HD-LBP	91.47	92.18	92.57
	HD-Gabor	90.73	92.33	92.80
Unrestricted	HD-LBP	93.43	94.25	94.87
	HD-Gabor	93.58	94.73	95.25

Table 1 shows the recognition accuracy on both restricted and unrestricted settings with different features. It is shown that with invisible region filling, the accuracy is improved by 1.38% with Gabor and 0.77% with LBP averagely. Considering there are only 3323 of 6000 testing pairs need invisible region filling, the improvement is significant. Besides, if we further preserve the background, there will be a stable 0.5% improvement for each feature and classifier, indicating that the external face region takes identity information helpful for face recognition.

5.1.2 Results and Discussions

In this part, we evaluate the performance of the proposed method following image-restricted protocol. Table 2 shows the face recognition accuracy of different methods and Fig. 9 shows the corresponding ROC curves.

We firstly list the results with the unsupervised learning PCA and demonstrate a 3.67% improvement by HPEN. Note that both as 3D normalization methods followed by high-dim feature and PCA, the HPEN outperforms the PAF by 1.38% since we explicitly normalize the face appearance. By applying DDML, the HPEN help improve the performance of HD-LBP and HD-Gabor by 1.79% and 1.85%, respectively. Although DDML can effectively learn discriminative metric [26], the HPEN preprocessing is able to further enhance the face recognition performance by simplifying the learning task with normalization.

Table 2. Mean classification accuracy and standard error on LFW under restricted, label-free outside data protocol.

Methods	Accuracy ($\bar{\mu} \pm S_E$)
PAF [48]	0.8777 ± 0.0051
Convolutional DBN [27]	0.8777 ± 0.0062
Sub-SML [15]	0.8973 ± 0.0038
DDML + Fusion [26]	0.9068 ± 0.0141
VMRS [7]	0.9110 ± 0.0059
HD-Gabor + PCA (Ours)	0.8548 ± 0.0032
HD-LBP + DDML (Ours)	0.9078 ± 0.0048
HD-Gabor + DDML (Ours)	0.9095 ± 0.0040
HPEN + HD-Gabor + PCA	0.8915 ± 0.0033
HPEN + HD-LBP + DDML	0.9257 ± 0.0036
HPEN + HD-Gabor + DDML	0.9280 ± 0.0047

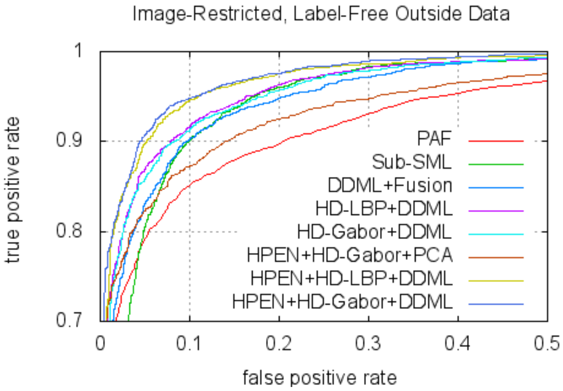

Figure 9. ROC curves under the LFW restricted, label-free outside data protocol.

We further examine the effectiveness of HPEN in unconstrained setting. Table 3 shows the mean accuracy and Fig. 10 shows the corresponding ROC curves.

Table 3. Mean classification accuracy and standard error on LFW under unrestricted, label-free outside data protocol.

Methods	Accuracy ($\bar{\mu} \pm S_E$)
Joint Bayesian [20]	0.9090 ± 0.0148
ConvNet-RBM [45]	0.9175 ± 0.0048
High-dim LBP [22]	0.9318 ± 0.0107
Aurora [33]	0.9324 ± 0.0044
FCN [51]	0.9438
HD-LBP + JB (Ours)	0.9347 ± 0.0059
HD-Gabor + JB (Ours)	0.9322 ± 0.0043
HPEN + HD-LBP + JB	0.9487 ± 0.0038
HPEN + HD-Gabor + JB	0.9525 ± 0.0036

The results show that HPEN improves the recognition results by 1.4% and 2.03% with HD-LBP and HD-Gabor respectively, where the the combination of HPEN, HD-Gabor and joint Bayesian reaches the state-of-the-art in unrestricted setting. Besides, the performance of HPEN is competitive with the facial component deep network (FCN), which is also an image level normalization method. Different from FCN, the proposed HPEN is learning-free, therefore good generalization performance can be expected.

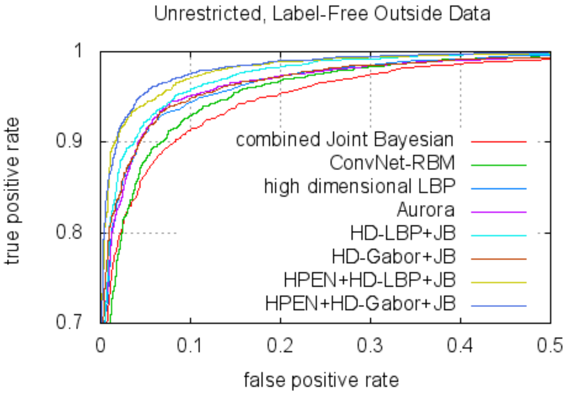

Figure 10. ROC curves under the LFW unrestricted, label-free outside data protocol.

5.2. Multi-PIE

The CMU Multi-PIE Face Database (MultiPIE) [25] contains images of 337 subjects collected under controlled environment with variations in pose, illumination and expression. Since Multi-PIE is highly organized and most of normalization methods are reported on this database, we can further analyze the robustness of HPEN to pose and expression. We conduct experiments on Multi-PIE with two setttings: **Setting-1** [50, 34, 5, 29] concentrates on pose variations. It uses images from all the 337 subjects at 7 poses $(-45°, -30°, -15°, 0°, 15°, 30°, 45°)$, with neural expression and frontal illumination (marked as 07) in all 4 sessions. The first 200 subjects are used as training set and the rest 137 subjects are used as testing set. During evaluation, the frontal images of each subject from the earliest session are used as gallery images and all remaining images are used as probe images. To further evaluate the robustness to simultaneous pose and expression variations, we propose the **Setting-2** which contains all the expressions including neutral, smile, surprise, squint, disgust and scream under poses of $0°$, $-15°$, $-30°$ and $-45°$ in frontal illumination. Other configurations are the same as Setting-1. This protocol is a extended and modified version of [23]. For each setting, the rank-1 recognition rates are reported, compared with the state-of-the-art methods.

In Setting-1, we demonstrate the robustness of our method to pose variations. Table 5 shows the comparison results with different normalization methods including a 3D method of Asthna11 [5] and four 2D methods of MDF [34], LE [17], FIP [50] and SPAE [29], all methods are conducted automatically. In this part, we sort methods into three

Table 4. Rank-1 recognition rates for MultiPIE in Setting-2 with simultaneous pose and expression variations, the results in the brackets are the recognition results without normalization. The expression "Smile" contains samples from both session 1 and session 3.

Expression/ Pose	Smile	Surprise	Squint	Disgust	Scream	Avg
05_1(0°)	99.31 (97.24)	98.44 (98.44)	98.44 (95.31)	95.83 (93.75)	89.01 (84.62)	96.21 (93.87)
14_0(−15°)	98.62 (97.93)	98.44 (95.31)	98.44 (95.31)	95.83 (92.17)	90.11 (80.22)	96.29 (92.30)
13_0(−30°)	96.55 (95.86)	95.31 (93.75)	95.31 (90.63)	94.79 (91.67)	83.52 (72.53)	93.10 (88.89)
08_0(−45°)	93.79 (93.10)	84.38 (79.69)	95.31 (92.19)	85.42 (87.50)	70.33 (61.54)	85.85 (82.80)
Avg	97.07 (96.03)	94.14 (91.80)	96.88 (93.36)	92.97 (91.41)	83.24 (74.73)	92.86 (89.46)

level of database dependence according to the data assumption they used. SPAE explicitly make the assumption that poses are sparse and discrete, thus it is of strong database dependence and as a result has difficulty in generalizing to unconstrained environment. Asthna11, LE and FIP do not utilize the data configuration, but their normalization models are trained on the same database with the testing set. Namely these methods make the assumption that the testing set shares the same pose variations with the training set, thus they have weak database dependence. MDF and HPEN do not have any assumption about the testing set, thus are database independent. In the experiment, we adopt the high-dim Gabor feature [22] as the feature extractor, and for better comparison we list the recognition results with both supervised classifier (LDA) which corresponds to SPAE, FIP, MDF, LE and unsupervised classifier (PCA) which corresponds to Asthna11.

Table 5. Rank-1 recognition rates for Multi-PIE in Setting-1, with the first and the second highest rates highlighted. The last column represents the database dependence, where "⋆⋆" means strong dependence, "⋆" means weak dependence and "-" means non dependence

Methods		Pose							Dep
		$-45°$	$-30°$	$-15°$	$15°$	$30°$	$45°$	*avg*	
2D	SPAE [29]	84.9	92.6	96.3	95.7	94.3	84.4	91.4	⋆⋆
	LE [17][1]	86.9	95.5	**99.9**	**99.7**	95.5	81.8	93.2	⋆
	FIP [50]	**95.6**	**98.5**	**100.0**	99.3	**98.5**	**97.8**	**98.3**	⋆
	MDF [34]	84.7	95.0	99.3	99.0	92.9	85.2	92.7	-
3D	Asthna11 [5]	74.1	91.0	95.7	95.7	89.5	74.8	86.8	⋆
	HPEN+PCA	88.5	95.4	97.2	98.0	95.7	89.0	94.0	-
	HPEN+LDA	**97.4**	**99.5**	99.5	**99.7**	**99.0**	**96.7**	**98.6**	-

In this setting, the HPEN demonstrates competitive results especially for large poses (±45°). Among geometry based normalization method, the HPEN outperforms the 3D Asthna11 and 2D MDF by 7.2% and 5.9% respectively, which may come from the good treatment for the invisible region and background. Compared with appearance transformation methods SPAE and FIP, HPEN also demonstrates competitive results and is believed to have better generalization ability due to the database independence. The improvements from the HPEN is demonstrated in Table 6.

Table 6. The average rank-1 recognition rates across poses (from $-45°$ to $45°$ except $0°$) on the original images and the normalized images with unsupervised and supervised classifiers.

Classifier	Original	Normalized	Error Reduced
PCA	86.5	94.0	55.5%
LDA	97.0	98.6	53.3%

In Setting-2, we evaluate the robustness of our method to simultaneous pose and expression variations. Table 4 shows the recognition results on both normalized and original images to demonstrate the improvement from our method. With HPEN, the average error of all the expressions and poses is reduced by 32.26%. However, the performance of HPEN deteriorates greatly when pose and expression are both far from the neutral condition, such as surprise and scream in $-45°$. The main reason is that landmark detector always fails in extreme conditions where many landmarks are too close and some are even invisible, which leads to inaccurate 3DMM fitting and bad normalization results.

6. Conclusion

In this paper, we propose a learning-free High-Fidelity Pose and Expression Normalization (HPEN) algorithm which could recover canonical-view, expression-free images of good quality. With HPEN, state-of-the-art performance is achieved in both constrained and unconstrained environments. However, there exist disadvantages of our method. Since HPEN fills the invisible region based on facial symmetry. If faces are occluded, the occluded region will be also mirrored, leading bad normalization results. This drawback will be improved in our future work.

7. Acknowledgment

This work was supported by the Chinese National Natural Science Foundation Projects #61203267, #61375037,

[1] The results come from [50]

#61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No.KGZD-EW-102-2, and AuthenMetric R&D Funds.

References

[1] O. Aldrian and W. A. Smith. Inverse rendering of faces with a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1080–1093, 2013.

[2] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[3] S. R. Arashloo and J. Kittler. Pose-invariant face matching using mrf energy minimization framework. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 56–69. Springer, 2009.

[4] A. B. Ashraf, S. Lucey, and T. Chen. Learning patch correspondences for improved viewpoint invariant face recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[5] 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 *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 937–944. IEEE, 2011.

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE, 2013.

[7] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz. Fast high dimensional vector multiplication face recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1960–1967. IEEE, 2013.

[8] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 545–552. IEEE, 2011.

[9] T. Berg and P. N. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *BMVC*, volume 2, page 7. Citeseer, 2012.

[10] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1063–1074, 2003.

[11] A. M. Bruckstein, R. J. Holt, T. S. Huang, and A. N. Netravali. Optimum fiducials under weak perspective projection. *International Journal of Computer Vision*, 35(3):223–244, 1999.

[12] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43, 2014.

[13] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Trans. Graph.*, 32(4):41, 2013.

[14] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: a 3d facial expression database for visual computing. 2013.

[15] Q. Cao, Y. Ying, and P. Li. Similarity metric learning for face recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2408–2415. IEEE, 2013.

[16] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2887–2894. IEEE, 2012.

[17] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2707–2714. IEEE, 2010.

[18] X. Chai, S. Shan, X. Chen, and W. Gao. Locally linear regression for pose-invariant face recognition. *Image Processing, IEEE Transactions on*, 16(7):1716–1725, 2007.

[19] C. H. Chan, M. A. Tahir, J. Kittler, and M. Pietikainen. Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1164–1177, 2013.

[20] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*, pages 566–579. Springer, 2012.

[21] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032. IEEE, 2013.

[22] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032. IEEE, 2013.

[23] B. Chu, S. Romdhani, and L. Chen. 3d-aided face recognition robust to expression and pose variations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1899–1906, 2014.

[24] L. Ding, X. Ding, and C. Fang. Continuous pose normalization for pose-robust face recognition. *Signal Processing Letters, IEEE*, 19(11):721–724, 2012.

[25] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[26] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1875–1882. IEEE, 2014.

[27] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2518–2525. IEEE, 2012.

[28] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2007.

[29] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1883–1890, 2014.

[30] A. Lapedriza, D. Masip, and J. Vitria. Are external face features useful for automatic face classification? In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 151–151. IEEE, 2005.

[31] Y. J. Lee, S. J. Lee, K. R. Park, J. Jo, and J. Kim. Single view-based 3d face reconstruction robust to self-occlusion. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–20, 2012.

[32] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3499–3506. IEEE, 2013.

[33] P. Li. Aurora face recognition technical report: Evaluation of algorithm aurora-c-2014-1 on labeled faces in the wild.

[34] 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 *Computer Vision–ECCV 2012*, pages 102–115. Springer, 2012.

[35] S. Lucey and T. Chen. A viewpoint invariant, sparsely registered, patch based, face verifier. *International Journal of Computer Vision*, 80(1):58–71, 2008.

[36] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 296–301. IEEE, 2009.

[37] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 313–318. ACM, 2003.

[38] U. Prabhu, J. Heo, and M. Savvides. Unconstrained pose-invariant face recognition using 3d generic elastic models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):1952–1961, 2011.

[39] S. J. Prince, J. Warrell, J. Elder, and F. M. Felisberti. Tied factor analysis for face recognition across large pose differences. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):970–984, 2008.

[40] C. Qu, E. Monari, T. Schuchert, and J. Beyerer. Fast, robust and automatic 3d face model reconstruction from videos. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 113–118. IEEE, 2014.

[41] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *Proc. BMVC*, volume 1, page 7, 2013.

[42] R. Singh, M. Vatsa, A. Ross, and A. Noore. A mosaicing scheme for pose-invariant face recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(5):1212–1225, 2007.

[43] L. Spreeuwers. Fast and accurate 3d face recognition. *International Journal of Computer Vision*, 93(3):389–414, 2011.

[44] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2013.

[45] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1489–1496. IEEE, 2013.

[46] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2013.

[47] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. CVPR, 2013.

[48] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3539–3545. IEEE, 2013.

[49] L. Zhang and D. Samaras. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):351–363, 2006.

[50] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 113–120. IEEE, 2013.

[51] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*, 2014.