Hierarchical-PEP Model for Real-world Face Recognition

Haoxiang Li, Gang Hua
Stevens Institute of Technology
Hoboken, NJ 07030
{hli18, ghua}@stevens.edu

Abstract

Pose variation remains one of the major factors adversely affect the accuracy of real-world face recognition systems. Inspired by the recently proposed probabilistic elastic part (PEP) model and the success of the deep hierarchical architecture in a number of visual tasks, we propose the Hierarchical-PEP model to approach the unconstrained face recognition problem. We apply the PEP model hierarchically to decompose a face image into face parts at different levels of details to build pose-invariant part-based face representations. Following the hierarchy from bottom-up, we stack the face part representations at each layer, discriminatively reduce its dimensionality, and hence aggregate the face part representations layer-by-layer to build a compact and invariant face representation. The Hierarchical-PEP model exploits the fine-grained structures of the face parts at different levels of details to address the pose variations. It is also guided by supervised information in constructing the face part/face representations. We empirically verify the Hierarchical-PEP model on two public benchmarks (i.e., the LFW and YouTube Faces) and a face recognition challenge (i.e., the PaSC grand challenge) for image-based and video-based face verification. The state-of-the-art performance demonstrates the potential of our method.

1. Introduction

In real-world face recognition, the difficulty comes from all kinds of visual variations including changes in expression, illumination, pose, and etc. Pose variation is one of the major challenges among these. The same face in different poses can look drastically different to each other as shown in Figure 1, Early works by Turk *et al.* [47] and Belhumeur *et al.* [4] in this area focus on recognizing well aligned frontal faces. They empirically demonstrate that frontal faces can be projected to a low-dimensional subspace invariant to variation in illumination and facial expressions [4]. This observation highlights the importance of addressing pose variation because it can greatly help relieve the adverse effects of the other visual variations.

Figure 1. Pose variation: the same person looks different in varied poses.

A line of research approaches this problem by generating the same-view face images given a face pair in presence of pose variations. For example, Prabhu *et al.* [40] use 3D face models to rotate the face image to an unseen view. Zhu *et al.* [54] recover the canonical-view faces with deep neural networks directly trained to predict the transformation from multi-view face images to their canonical views. In these methods, they try to holistically align faces to relieve the pose variation.

Another set of methods resort to locate facial landmarks to build pose-invariant face representations [10, 15]. For example, Chen *et al.* [15] concatenate dense features around the facial landmarks to build the face representation. The pose-invariance is achieved in this way, because it always extracts features from the face part surrounded around the facial landmarks regardless of their locations in the image.

The elastic matching methods [20, 29, 30, 50] generalize this design. Li *et al.* [29, 30] present a probabilistic elastic part (PEP) model unsupervisedly learned from face image patches. The PEP model is a mixture of part models, each of which implicitly defines a face part. The PEP model looks for the image patches from these face parts for faces in varied poses. It then constructs the face representation by concatenating features from these selected image patches.

This procedure – locating the face parts and stacking the features to build face representation – is demonstrated to be effective by both Chen *et al.* [15] and Li *et al.* [29]. In extracting feature of the face part, Chen *et al.* [15] use the high-dimensional stacked SIFT [33] features and Li *et al.* [29] simply use the SIFT feature. Although low-level

features like SIFT present partial invariance to local variations, we argue that directly describing the face parts with naive dense extraction of these low-level features may not be optimal.

In this work, we propose to build a better face part model to construct an improved face representation. We base our method on the PEP model [29] to construct probabilistic elastic part based representation for the face parts. In our method, we model the face parts at different levels of details in a hierarchical fashion. And we build the face part representation by locating subtle structures of the face part and stacking the features from different levels together. In this way, we construct the pose-invariant face representation in a hierarchical structure. We call the new model to be a Hierarchical-PEP model and the new *parts of parts* face representation to be POP-PEP.

The hierarchical structure may produce a very high dimensional face representation. We avoid this drawback by discriminative dimensionality reduction in constructing the face part representation. Moreover, this dimensionality reduction is applied from the bottom-level up to the holistic face with a simple network of Principle Component Analysis (PCA) and Linear Discriminant Embedding (LDE) [16]. The similar technique has been adopted by Simonyan *et al.* [42] and Chan *et al.* [12]. Chan *et al.* [12] present a simple deep network with cascaded PCA. Simonyan *et al.* [42] iteratively conduct PCA and spatial stacking to form a simple deep networks.

In this work, we further integrate the supervised information in a similar structure in aggregating the part representations. We iteratively stack the representations of the subtle face part structures and apply the discriminative dimensionality reduction. We empirically validated the effectiveness of this design in our experiments (see Section 4 for details).

Our contributions in this work are three-fold:

- we present a Hierarchical-PEP model to exploit the subtle face part structures at different levels of details for improving pose invariance;
- we propose a simple network of discriminative dimensionality reduction to integrate the face part representations to a compact and discriminative face representation;
- we achieve state-of-the-art performance on two public face verification benchmarks and a face recognition challenge.

2. Related Work

Face recognition has been an active research topic for tens of years. Recently, the well-designed challenging face recognition benchmarks [24, 48] and emerging face technology applications foster the development of a number of real-world face recognition methods [23, 3, 5, 6, 9, 13, 18, 28, 32, 34, 35, 44, 46].

To construct pose-invariant face representation, previous work has proposed to explicitly address the pose variations with 3D information. For example, Prabhu *et al.* [40] use 3D face models to rotate the gallery face image to the estimated view of the probe image; Yi *et al.* [51] use a 3D deformable model to estimate the face pose and apply the pose adaptive filters for feature extraction; Li *et al.* [31] propose to learn the morphable displacement fields from 3D face models to synthesize the probe face in the same view of the gallery face.

With only the 2D information, Yin *et al.* [52] propose the associate-predict model to transfer appearance of an alike identity in the database to approximate the appearance of the probe face at an unseen pose; Chen *et al.* [15] extract features densely at the facial landmarks and stack the features as the high-dimensional face representation.

The most relevant work to ours is the PEP model [29]. The PEP model consists of a set of face part models obtained through unsupervised learning. Given a face image, each face part model selects the most similar image patch. The PEP model then achieves the pose invariance by concatenating features extracted from the selected image patches to represent the face image.

In this work, the Hierarchical-PEP model hierarchically exploits the face parts and discriminatively integrate the part representations. Besides producing more discriminative face representation, the Hierarchical-PEP model shares the advantages of the PEP model that it builds representation for both face images and face videos in an unified framework and it does not require massive training data.

Besides the methods based on conventional hand-crafted features, a number of deep learning methods including the DeepID [44], DeepFace [46] and Stacked Progressive Auto-Encoders [25] are successfully applied to the face recognition problem, which achieve significantly improved verification accuracy. Despite the highly accurate recognition rate, these systems require a large number of labeled data in the training stage.

In this work, we focus on understanding the face recognition problem without leveraging massive training data. And we observe that the face recognition system based on the conventional feature descriptors can benefit from a hierarchical structure. The similar observation is reported by Simonyan *et al.* [42] in the general image classification task in which they present a 2-layer fisher vector encoding based networks based on the SIFT feature.

3. Hierarchical-PEP model

3.1. Introduction to the PEP model

The Hierarchical-PEP model consists of a hierarchy of PEP models. Formally, the PEP model [29] is denoted and parameterized as

$$\mathcal{P}(L, \{\mathcal{G}_k\}_{k=1}^{K}) \quad (1)$$

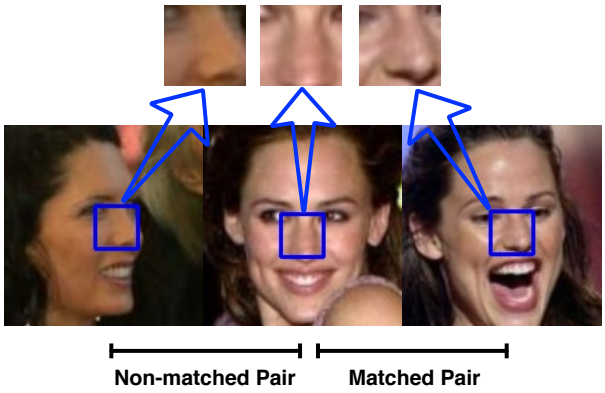

Figure 2. Image patches from the same face part are varied in visual appearance.

where $\mathcal{P}$ is the PEP model of K mixture components. Each mixture component is a face part model $\mathcal{G}_k$. L is the size of the face part. Given a testing image, the face part model $\mathcal{G}_k$ identifies the face part of size $L \times L$. More specifically, $\mathcal{G}_k$ is a spherical Gaussian model. Given a face image divided as N $L \times L$ image patches $\{I_n\}_{n=1}^N$, $\mathcal{G}_k$ chooses $I_{n^\star}$ with the highest probability. Formally,

$$p_n = [\mathbf{a}_n \, \mathbf{l}_n], \tag{2}$$

$$\mathcal{G}_k = \mathcal{N}(p_n|\vec{\mu}_k, \sigma_k^2\mathbf{I}), \tag{3}$$

where p_n is the representation of the image patch I_n; $\mathbf{a}_n$ is the appearance feature descriptor extracted from I_n (*e.g.* a SIFT descriptor); $\mathbf{l}_n$ is the spatial location of the image patch I_n in the full face image; $\mathbf{I}$ is an identity matrix; $\vec{\mu}_k$ and σ_k^2 are the mean and variance of the Gaussian model respectively [1]. $\mathcal{G}_k$ selects $I_{n_k^\star}$ that

$$n_k^\star = \arg\max_n \mathcal{N}(p_n|\vec{\mu}_k, \sigma_k^2\mathbf{I}). \tag{4}$$

Given a face image f, the PEP model produces the face representation $\mathcal{F}_\mathcal{P}(f) = \mathcal{B}(\mathcal{P}, f)$, where $\mathcal{B}$ denotes the representation construction process. Specifically, the k-th face part model $\mathcal{G}_k$ produces the face representation $\mathcal{F}_{\mathcal{G}_k}(f) = \mathcal{B}(\mathcal{G}_k, f)$,

$$f = \{p_n\}_{n=1}^N, \tag{5}$$

$$\mathcal{B}(\mathcal{G}_k, f) = \mathbf{a}_{n_k^\star}, \tag{6}$$

$$\mathcal{B}(\mathcal{P}, f) = [\,\{\mathcal{B}(\mathcal{G}_k, f)\}_{k=1}^K], \tag{7}$$

where $n_k^\star$ indicates the feature descriptor of the image patch identified by the face part model $\mathcal{G}_k$ as in Equation 4. The PEP model $\mathcal{P}$ then builds the face representation $\mathcal{F}_\mathcal{P}(f)$ as the concatenation of $\mathcal{F}_{\mathcal{G}_k}(f)$, $k = 1 \ldots K$.

One of the advantages of the PEP model is that it processes image and video in an unified framework. Given an M-frame face video $v = \{f_m\}_{m=1}^M$, the PEP model builds the video face representation $\mathcal{F}_\mathcal{P}(v) = \mathcal{F}_\mathcal{P}(\cup_{m=1}^M f_m)$. For brevity, we only take face images as examples in the following sections. The video face can be processed in the same framework.

[1] The covariance matrix is restricted to be spherical to mix the constraint from the appearance feature and spatial location to balance the influence from the two parts, as advocated in [29].

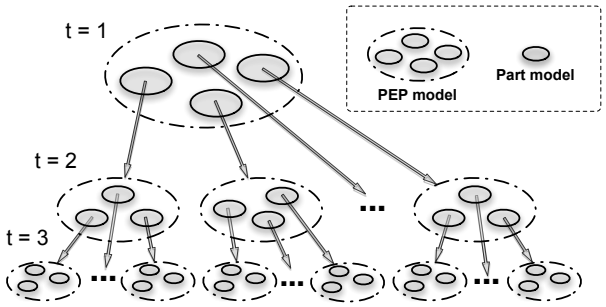

Figure 3. Example 3-layer Hierarchical-PEP model: a hierarchy of PEP models.

We refer readers to [29] for the detailed training process of the PEP model. In brief, to obtain a PEP model parameterized by L and K, the training face images are first processed into $L \times L$ densely sampled image patches; appearance descriptors are extracted from the image patches and concatenated with the spatial locations of the patches in the full face image; a K-component Gaussian mixture model is then learned through the Expectation-Maximization (EM) algorithm from the training feature descriptors to be the PEP model.

3.2. Hierarchical-PEP model

The effectiveness of the PEP model is originated from its capability to locate the face parts. Previous works [29, 30] have empirically shown the PEP model builds pose-invariant representations for faces. Given a face f, the PEP model builds its representation as the concatenation of a sequence of appearance descriptors. However, describing the face parts with the low-level feature descriptor (*e.g.*SIFT) can be suboptimal for face verification.

As the example shown in Figure 2, although the same face parts are correctly identified from the three face images, the pose change still adversely affects the matching of the selected patches. Motivated by this observation, we propose to further apply another PEP model at a more fine-grained level, *i.e.*, with smaller image patch size L, to match the image patches identified by the face part models. Instead of extracting general low-level descriptor to describe the face part, we build another layer of pose-invariant face part PEP representations to describe each face part at the previous level.

A T-layer Hierarchical-PEP model is shown in Figure 3 ($T = 3$). A Hierarchical-PEP $\mathcal{H}_t$ at layer t consists of

1. the PEP model $\mathcal{P}_t$ of K_t mixture components operating on face parts in size $L_t \times L_t$;

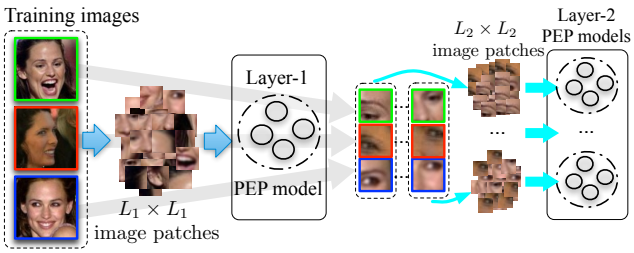

Figure 4. Example training process of a 2-layer Hierarchical-PEP model: image patches from the same face are in the same color.

2. K_t Hierarchical-PEP models $\{\mathcal{H}_{t+1}^k\}_{k=1}^{K_t}$ at layer $t+1$ if $t < T$.

3.2.1 Training of the Hierarchical-PEP model

The training process of the Hierarchical-PEP model is illustrated in Figure 4. Given a set of training face images $F = \{f_i\}_{i=1}^{|F|}$, we recursively train the T-layer Hierarchical-PEP model. We first learn a PEP model $\mathcal{P}$ from F. Following the Equation 4, the k-th face part model processes all $|F|$ training face images and identifies $|F|$ image patches from F. The set of identified image patches by the k-th face part model is denoted as F_k. Then we follow the same process to train a $(T-1)$-layer Hierarchical-PEP model from F_k.

3.2.2 Top-down image patch selection

As shown in Figure 5, given an image I, we locate the face parts and the sub-parts with the Hierarchical-PEP model following the top-down work-flow.

For the Hierarchical-PEP model $\mathcal{H}_t$, the input image is denoted as I_t. the image is processed into a set of $L_t \times L_t$ patches and K_t image patches are identified by the K_t face part models in the PEP model $\mathcal{P}_t$ following the Equation 4. The k-th face part model identifies the image patch I_{t+1}^k. If $t < T$, the Hierarchical-PEP model $\mathcal{H}_{t+1}^k$ further processes the image I_{t+1}^k. The input image for the full Hierarchical-PEP model $\mathcal{H}_1$ is the full face image I.

A sample top-down image patch selection result for a face pair along a path in the hierarchy is shown in Figure 6. We can observe the elastic matching effects at different levels of details.

3.2.3 Bottom-up representation construction

We follow the top-down image patch selection work-flow to obtain the input images for all the Hierarchical-PEP models and then start from the bottom to aggregate the representations.

Following the previous notations for the face representation, a Hierarchical-PEP at layer t can build representation

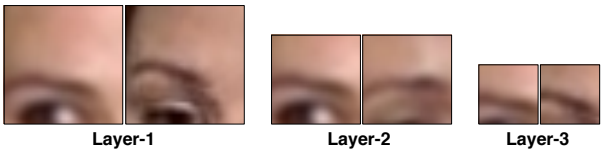

Figure 6. Example face parts selected in the top-down image patch selection process: the leftmost pair is two patches selected from the full face image by a face part model in the layer-1 of the Hierarchical-PEP; the pair in the middle is selected from a sub-structure of this face part; the rightmost pair describes a more subtle structure. We can observe the matching becomes more fine-grained and accurate in the later layers.

$\mathcal{B}(\mathcal{H}_t, I_t)$ given the input image I_t,

$$\mathcal{B}(\mathcal{H}_t, I_t) = \begin{cases} [\,\{\mathcal{B}(\mathcal{H}_{t+1}^k, I_{t+1}^k)\}_{k=1}^{K_t}\,], & \text{if } t < T, \\ \mathcal{B}(\mathcal{P}_t, I_t), & \text{if } t = T. \end{cases} \tag{8}$$

As shown in Figure 5 (ignoring the PCA/LDE legends introduced later), in layer-2 the Hierarchical-PEP models build the PEP representations (stacked SIFT descriptors describing the face parts) and the representations are then stacked as the upper layer representations to represent the full face image.

3.2.4 Discriminative dimensionality reduction

Following the above bottom-up representation construction process, we build a face representation in the form of a $\Pi_{t=1}^T K_t \times D$ dimensional vector, where D is the dimensionality of the chosen appearance feature descriptor, *i.e.*, D equals 128 for SIFT descriptor. This representation could be of very high dimensionality in practice. Hence it is highly favorable to reduce its dimensionality for efficient storage and computation.

Given a set of matched and non-matched training face pairs, Li *et al.* [30] propose to apply PCA to reduce the dimensionality of the PEP representation and utilize the Joint Bayesian classifier [14] for verification. The same process is applicable to the Hierarchical-PEP representation. However, we prefer to have a discriminative face representation while the Joint Bayesian classifier produces a discriminative similarity measurement. We resort to the Linear Discriminant Embedding (LDE) [16] method to find a subspace with smaller intra-class (matched faces) variations and larger inter-class (non-matched faces) variations.

We first reduce the dimensionality of the Hierarchical-PEP representations by PCA. We then look for the subspace which enlarges the distance between non-matched face pairs and shrinks the distance between matched face pairs, *i.e.*,

$$\bar{\mathbf{w}} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}}, \tag{9}$$

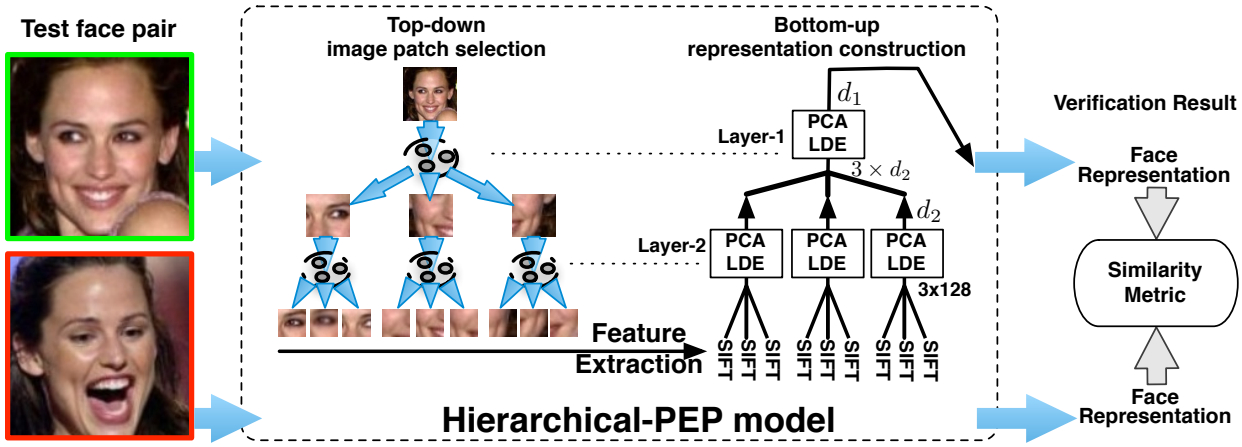

Figure 5. Construction of the face representation with an example 2-layer Hierarchical-PEP model: PCA at layer t keeps d_t dimensions.

where

$$\begin{aligned}
\mathcal{F}(I_i) &= \mathbf{PCA}(\mathcal{B}(\mathcal{H}, I_i)), \\
\mathcal{F}(I_j) &= \mathbf{PCA}(\mathcal{B}(\mathcal{H}, I_j)), \\
\mathbf{A} &= \sum_{l_{ij}=0} (\mathcal{F}(I_i) - \mathcal{F}(I_j))(\mathcal{F}(I_i) - \mathcal{F}(I_j))^T, \\
\mathbf{B} &= \sum_{l_{ij}=1} (\mathcal{F}(I_i) - \mathcal{F}(I_j))(\mathcal{F}(I_i) - \mathcal{F}(I_j))^T,
\end{aligned}$$

where the PCA dimensionality reduction process is denoted as $\mathbf{PCA}$; $l_{ij} = 1$ when $I_{i,j}$ is a matched face pair, $l_{ij} = 0$ otherwise. With the LDE projection, for one face image we can obtain a low-dimensional discriminative representation for face verification.

3.2.5 Integration of supervised information

The PCA and LDE projections discriminatively reduce the dimensionality of the face representation. The same process is also applicable in constructing representations at the face part level.

Recursively from the bottom-up, we stack the part representations and apply the PCA and LDE to construct the upper-level representation. With the simple network of PCA/LDE shown in Figure 5, the supervised information is hierarchically integrated into the face representation.

In the integration process shown in Figure 5, we apply the PCA and LDE to the PEP representations built at the bottom layer of the Hierarchical-PEP model. Then instead of aggregating the PEP representations, we aggregate the low-dimensional discriminative representations to the upper layer. In this way, all the aggregated representations in the hierarchy are not only pose-invariant but also discriminative for face verification.

The Equation 8 is updated for this process as

$$\mathcal{B}(\mathcal{H}_t, I_t) = \begin{cases} \mathbf{DR}([\,\{\mathcal{B}(\mathcal{H}_{t+1}^k, I_{t+1}^k)\}_{k=1}^{K_t}\,]), & \text{if } t < T, \\ \mathbf{DR}(\mathcal{B}(\mathcal{P}_t, I_t)), & \text{if } t = T, \end{cases} \tag{10}$$

where $\mathbf{DR}(X) = \mathbf{LDE}(\mathbf{PCA}(X))$.

3.3. Hierarchical-PEP model for Face Verification

The Hierarchical-PEP model builds the discriminative low-dimensional POP-PEP face representation. In the face verification task, given two face images I_1 and I_2 the T-layer Hierarchical-PEP model constructs the face representations $\mathcal{B}(\mathcal{H}_1, I_1)$ and $\mathcal{B}(\mathcal{H}_1, I_2)$ following Equation 10. The similarity score of the two faces is simply the cosine similarity (or dot-product after normalization) of the two face representations

$$s(I_1, I_2) = \frac{\mathcal{B}(\mathcal{H}_1, I_1)\mathcal{B}(\mathcal{H}_1, I_2)}{|\mathcal{B}(\mathcal{H}_1, I_1)||\mathcal{B}(\mathcal{H}_1, I_2)|}. \tag{11}$$

3.3.1 Multiple layers fusion

Given a T-layer Hierarchical-PEP model we can truncate all the leaf PEP models to get a $(T-1)$-layer Hierarchical-PEP model. Given the $(T-1)$-layer Hierarchical-PEP model, it constructs the face representation at a different level of detail. The observation in the previous work adopting a coarse-to-fine structure [17, 27] suggests that we can benefit from fusing the representations or the scores across the coarse-to-fine structure.

That is, in building the face representation, as shown in Figure 5, we can follow the top-down work-flow to the last but one layer above and aggregate bottom-up from there to obtain the face representation. More specifically, given the face pair I_1 and I_2 we can set the T in Equation 10 to $t' < T$ to obtain the confidence score $s_{t'}$ with a t'-layer Hierarchical-PEP model. The final confidence score of the face pair is the average score $s(I_1, I_2) = \frac{1}{T}\sum_{t=1}^{T} s_t$. In our experiments, we observe consistent improvement with this multiple layers fusion.

4. Experimental Evaluation

We evaluate the Hierarchical-PEP model for both the image-based face verification and video-based face verification. Sharing the same advantages of the PEP model, the

Table 1. Performance comparison with the baseline methods.

Algorithm	Accuracy ± Error(%)
a) 1-layer, 4096-component	89.30 ± 1.33
b) 3-layer, w/o LDE	88.00 ± 1.80

Hierarchical-PEP model builds representations for face images and face videos in an unified framework.

4.1. Labeled Faces in the Wild

The Labeled Faces in the Wild (LFW) dataset [24] is designed as a benchmark for uncontrolled image-based face verification. This dataset contains $13,233$ images from $5,749$ people. LFW defines six protocols for fair comparison [22]. Without accessing outside training data, we train our face recognition system in the restricted setting. Specifically, we report the 10-fold average accuracy under the image-restricted with no outside data protocol. In our experiments, we set the parameters as a computation-accuracy trade off instead of just focusing on accuracy. Potential improvement may be obtained with more aggressive setting.

4.1.1 Settings

Following the predefined protocol on LFW, we use the images roughly aligned by the funneling algorithm [21]. We center crop the images to size 150×150 to exclude most of the background to focus on recognizing the face.

We train a 3-layer Hierarchical-PEP model ($T = 3$). The first layer consists of a PEP model with 256 face part models ($K_1 = 256$) working on image patches of size 32×32 ($L_1 = 32$). The second layer consists of PEP models with 4 face part models ($K_2 = 4$) working on image patches of size 24×24 ($L_2 = 24$). The last layer consists of PEP models with 4 face part models ($K_3 = 4$) working on image patches of size 16×16 ($L_3 = 16$). We set $d_1 = 200$, $d_2 = 100$, and $d_3 = 50$. The final face representation is of 200 dimensions. We keep other parameters consistent to the Eigen-PEP [30] model for fair comparison.

The parameters are chosen under the consideration to keep the computational expense acceptable on typical CPU workstations. It takes 41 hours to train the Hierarchical-PEP model including the PCA and LDE projections on a PC with 12 CPU cores [2].

4.1.2 Results

As shown in Figure 7 and Table 2, we observe that the Hierarchical-PEP model achieves very competitive accuracy. While Li *et al.* [30] combine the SIFT feature and Local Binary Patterns (LBP) [1] to obtain an average 88.97% accuracy, we achieve 91.10% accuracy with SIFT feature only. In Table 1, we further present some baseline results to explore how the steps in the proposed method contribute to the performance improvement.

[2]Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz

Table 2. Performance comparison on LFW under the Image-restricted with no outside data protocol.

Algorithm	Accuracy ± Error(%)
APEM [29]	84.08 ± 1.20
Fisher vector faces [41]	87.47 ± 1.49
Eigen-PEP [30]	88.97 ± 1.32
Hierarchical-PEP (1-layer)	87.20 ± 1.39
Hierarchical-PEP (2-layer)	90.00 ± 1.29
Hierarchical-PEP (3-layer)	90.40 ± 1.35
Hierarchical-PEP (layers fusion)	$\mathbf{91.10 \pm 1.47}$
Hierarchical-PEP (3-layer) with Joint Bayesian Classifier trained from labeled pairs	90.37 ± 1.22
Published after submission: MRF-Fusion-CSKDA[2]	$\mathbf{95.89 \pm 1.94}$

In the baseline experiment a), we compare with a 1-layer Hierarchical-PEP model with 4096 components ($L_1 = 16$, $d_1 = 200$). Without the hierarchy but keeping the total number of Gaussian components the same as the 3-layer Hierarchical-PEP model, the 10-fold average accuracy degrades. It demonstrates the hierarchical architecture helps improve the performance.

In the baseline experiment b), we remove the LDE process from the bottom-up aggregation but still keep the PCA for dimensionality reduction. We only apply the LDE for the final face representations instead. We observe that without the LDE in lower layers it underperforms the previous method. It demonstrates the effectiveness of the simple discriminative dimensionality reduction network in constructing the face representation.

Arashloo *et al.* [2] achieve a higher accuracy by fusing three kinds of descriptors MLBP [11], LPQ [45] and BSIF [26]. With only the MLBP descriptor, their accuracy is 90.68% while our result with solely SIFT is 91.1%. Besides, their method relies on the Markov Random Field to address the pose variation of which the optimization procedure is highly computationally expensive. In addition to LFW, we also evaluate our method on other two datasets (Section 4.2 and 4.3).

We also evaluate the Joint Bayesian classifier [14] with the POP-PEP face representation to compare with the results in [30]. Within the image-restricted protocol, we follow Li *et al.* [30] to train the Joint Bayesian classifier with labeled face pairs to learn similarity scores for face pairs. However, we observe no improvement compared with the simple cosine similarity. In Section 4.3.3, we conduct more experiments on the PaSC dataset to explore whether the Joint Bayesian classifier can improve the recognition accuracy with the Hierarchical-PEP face representation.

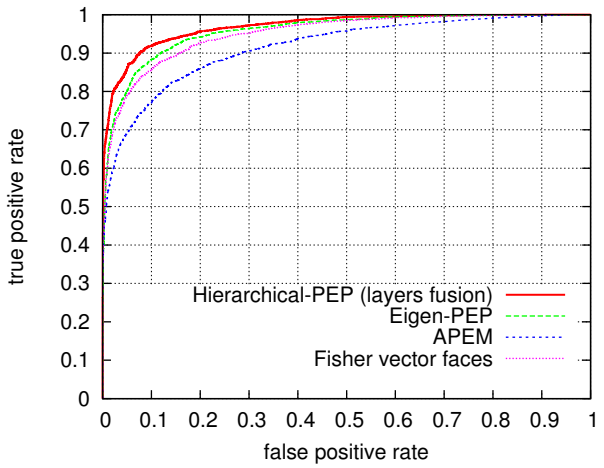

Figure 7. Performance comparison on LFW under the Image-restricted with no outside data protocol.

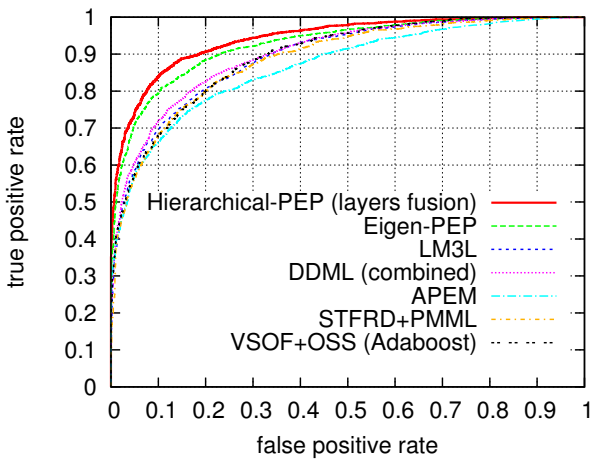

Figure 8. Performance comparison on YTF under the restricted with no outside data protocol.

Table 3. Performance comparison on YTF with different numbers of frames per video with the 2-layer Hierarchical-PEP model.

# frames	Accuracy $\pm$ Error(%)
10	85.40 ± 1.36
50	86.84 ± 1.35
all (181 on average)	87.00 ± 1.50

Table 4. Performance comparison on YTF under the restricted with no outside data protocol.

Algorithm	Accuracy $\pm$ Error(%)
MBGS [48]	76.4 ± 1.8
MBGS+SVM- [49]	78.9 ± 1.9
STFRD+PMML [53]	79.5 ± 2.5
VSOF+OSS(Adaboost) [38]	79.7 ± 1.8
APEM [29])	79.10 ± 1.50
VF^2 [39]	84.7 ± 1.4
DDML (combined) [18]	82.3 ± 1.5
Eigen-PEP [30]	84.8 ± 1.4
LM3L [19]	81.3 ± 1.2
Hierarchical-PEP (1-layer)	86.16 ± 1.11
Hierarchical-PEP (2-layer)	86.72 ± 1.51
Hierarchical-PEP (layers fusion)	$\mathbf{87.00 \pm 1.50}$

4.2. YouTube Faces

The YouTube Faces (YTF) dataset [48] follows the design of LFW as a benchmark for uncontrolled video-based face verification. This dataset contains $3,425$ videos from $1,595$ people. Each video consists of faces of the same person. On average, a video has 181 frames. We report our results under the image-restricted with no outside data protocol. On this dataset, the Hierarchical-PEP model further improves the state-of-the-art accuracy.

4.2.1 Settings

We center crop the video frames to be 100×100 to exclude most of the background and reduce computation. Considering the low resolution of the face videos, we train a 2-layer Hierarchical-PEP model ($T = 2$). The first layer consists of a PEP model with 256 face part models ($K_1 = 256$) working on image patches of size 32×32 ($L_1 = 32$). The second layer consists of PEP models with 16 face part models ($K_2 = 16$) working on image patches of size 24×24 ($L_2 = 24$). We set $d_1 = 400$ and $d_2 = 200$. The final face representation is of 400 dimensions. Other settings are the same as in Section 4.1.1.

4.2.2 Results

As shown in Figure 8 and Table 4, we observe the Hierarchical-PEP model significantly improves the state-of-the-art accuracy. The 2-layer Hierarchical-PEP model consistently improves the accuracy of a 1-layer model and the multiple layers fusion can further improve the accuracy.

In Table 3, we show how the performance improves by adding more frames in building the face video representations. We observe that with 10 frames randomly selected for each face video, the Hierarchical-PEP model achieves state-of-the-art performance.

4.3. Point-and-Shoot Face Recognition Challenge

Beveridge *et al.* [7] propose the Point-and-Shoot Face Recognition Challenge (PaSC) to facilitate the development of uncontrolled video face recognition algorithms. The PaSC includes $9,376$ still images of 293 people and $2,802$ videos for 265 people balanced with varied factors such as the distance to the camera, viewpoints, the sensor types and etc. We refer the readers to the report by Beveridge *et al.* [7] for more details.

There are two experiments defined in the PaSC, the video-to-video experiment and the video-to-still experi-

Table 5. Evaluation on the PaSC of cosine similarity and Joint Bayesian classifier trained with labeled pairs.

Experiment	layer-1	layer-2	fusion
exp1 (cosine)	0.199	0.206	0.212
exp1 (Joint Bayesian)	0.195	0.251	0.258
exp2 (cosine)	0.259	0.284	0.299
exp2 (Joint Bayesian)	0.226	0.264	0.288

ment. In the video-to-video experiment, given the target and query sets of videos respectively, the participants are asked to report the pairwise video-to-video similarities of the two sets and report the verification accuracy at the 0.01 false alarm rate. In the video-to-still experiment, the settings are the same except that the target set contains still images instead of videos. We evaluate our method on the PaSC dataset and compare with the results reported in [8].

4.3.1 Settings

We use the eye-coordinates provided by the PaSC organizers to align faces and crop out to 150×150 images for a fair comparison. Considering the low resolution of the videos, we train a 2-layer Hierarchical-PEP model ($T = 2$) on the LFW dataset. The first layer consists of a PEP model with 256 face part models ($K_1 = 256$) with patch size 32×32 ($L_1 = 32$). The second layer consists of PEP models with 16 face part models ($K_2 = 16$) working on image patches of size 24×24 ($L_2 = 24$). We set $d_1 = 100$ and $d_2 = 50$. The final face representation is of 100 dimensions. Other settings are the same as in Section 4.1.1.

We train this Hierarchical-PEP model with the $6,000$ pairs of face images in LFW roughly aligned by the funneling algorithm [21]. We then construct face representations for all the $13,233$ face images in LFW and train a Joint Bayesian classifier with their identity labels following Chen *et al.* [14].

4.3.2 Results

We report verification accuracy at 0.01 false alarm rate in the PaSC in the video-to-video experiment (exp1) and video-to-still experiment (exp2). Although the two datasets are very different, our method shows very nice generalization. When applying the system we trained on the LFW to the PaSC directly our system largely outperforms the best results in the two experiments, as shown in Table 6.

4.3.3 Joint Bayesian Classifier

The Joint Bayesian classifier models the extra-person and intra-person variations as zero-mean Gaussians with two different covariance matrices. Empirically, it outperforms the Linear Discriminant Analysis (LDA) [4] in face recognition. We refer readers to [14] for more details.

Table 6. Performance comparison on the PaSC.

Algorithm	exp1	exp2
LPB-SIFT-WPCA-SILD [37]	0.09	0.23
ISV-GMM [36]	0.05	0.11
PLDA-WPCA-LLR [43]	0.19	0.26
LRPCA Baseline [7]	0.08	0.10
Eigen-PEP [30]	0.26	0.24
Hierarchical-PEP(1-layer)	0.261	0.275
Hierarchical-PEP(2-layer)	0.287	0.289
Hierarchical-PEP(2 layers fusion)	**0.307**	**0.320**

On LFW we observe that the Joint Bayesian classifier trained from labeled face pairs performs comparable to the cosine similarity with the Hierarchical-PEP face representations. On PaSC, the experimental results further support this observation. We use the $6,000$ labeled face pairs in LFW for training and compare the cosine similarity with Joint Bayesian classifier in Table 5. The Joint Bayesian classifier outperforms the cosine similarity in the video-to-video experiment but underperforms the cosine similarity in the video-to-still experiment.

This observation suggests that the simple PCA and LDE networks can already exploit the supervised information to build robust face representations when only the labels for face pairs are available. With additional identity labels, the Hierarchical-PEP face representation can benefit from more discriminative similarity metric such as the scores from the Joint Bayesian classifier for further improvement.

5. Conclusion

We propose a Hierarchical-PEP model for real-world face recognition. From top-down, the Hierarchical-PEP model hierarchically builds pose-invariant face representation for both face images and face videos in an unified framework. The Hierarchical-PEP model builds pose-invariant representations for face parts and its fine-grained structures. The part-based representations are then aggregated from bottom-up to construct the face representation. Supervised information is integrated in the aggregation process through a simple discriminative dimension reduction network. The Hierarchical-PEP model finally constructs low-dimensional discriminative full face representation for face verification. We observe a simple multiple layers fusion method consistently improves the accuracy. We evaluate the Hierarchical-PEP model on the LFW, YTF and PaSC datasets for the image-to-image, video-to-video and video-to-image face verification. The state-of-the-art performance demonstrates the effectiveness of the Hierarchical-PEP model. How to speed up the computation to efficiently adopt a more aggressive and potential setting remains a question to be addressed in our future work.

Acknowledgment

Research reported in this publication was partly supported by the National Institute Of Nursing Research of the National Institutes of Health under Award Number R01NR015371. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work is also partly supported by US National Science Foundation Grant IIS 1350763 and GH's start-up funds from Stevens Institute of Technology.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *Proc. European Conference on Computer Vision*, 2004. 6

[2] S. Arashloo and J. Kittler. Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features. *Information Forensics and Security, IEEE Transactions on*, 2014. 6

[3] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz. Fast high dimensional vector multiplication face recognition. In *Proc. IEEE International Conference on Computer Vision*, 2013. 2

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997. 1, 8

[5] T. Berg and P. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *British Machine Vision Conference*, 2012. 2

[6] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013. 2

[7] J. Beveridge, P. Phillips, D. Bolme, B. Draper, G. Givens, Y. M. Lui, M. Teli, H. Zhang, W. Scruggs, K. Bowyer, P. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *BTAS*, 2013. 7, 8

[8] J. R. Beveridge, H. Zhang, P. J. Flynn, Y. Lee, V. E. Liong, J. Lu, M. de Assis Angeloni, T. de Freitas Pereira, H. Li, G. Hua, V. Struc, J. Krizaj, and P. J. Phillips. The ijcb 2014 pasc video face and person recognition competition. *International Joint Conference on Biometrics (IJCB).*, 2014. 8

[9] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun. A practical transfer learning algorithm for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013. 2

[10] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1

[11] C.-H. Chan, J. Kittler, and K. Messer. Multi-scale local binary pattern histograms for face recognition. In *Proceedings of the 2007 International Conference on Advances in Biometrics*, 2007. 6

[12] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. Pcanet: A simple deep learning baseline for image classification? *arXiv preprint arXiv:1404.3606*, 2014. 2

[13] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Computer Vision–ECCV 2014*. 2014. 2

[14] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Proc. European Conference on Computer Vision*, 2012. 4, 6, 8

[15] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High dimensional feature and its efficient compression for face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1, 2

[16] J. Duchene and S. Leclercq. An optimal transformation for discriminant and principal component analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1988. 2, 4

[17] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Proc. IEEE International Conference on Computer Vision*, 2005. 5

[18] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2, 7

[19] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *ACCV*, 2014. 7

[20] G. Hua and A. Akbarzadeh. A robust elastic and partial matching metric for face recognition. In *Proc. IEEE International Conference on Computer Vision*, 2009. 1

[21] G. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *Proc. IEEE International Conference on Computer Vision*, 2007. 6, 8

[22] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. 6

[23] 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*, 2012. 2

[24] 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. Technical report, University of Massachusetts, Amherst, 2007. 2, 6

[25] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[26] J. Kannala and E. Rahtu. Bsif: Binarized statistical image features. In *Pattern Recognition (ICPR), 21st International Conference on*, 2012. 6

[27] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, 2006. 5

[28] Z. Lei, M. Pietikainen, and S. Z. Li. Learning discriminant face descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 2

[29] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Proc.*

IEEE Conference on Computer Vision and Pattern Recognition, 2013. 1, 2, 3, 6, 7

[30] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-pep for video face recognition. In *ACCV*, 2014. 1, 3, 4, 6, 7, 8

[31] 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. European Conference on Computer Vision*, 2012. 2

[32] S. Liao, A. Jain, and S. Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 2

[33] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 2004. 1

[34] C. Lu and X. Tang. Learning the face prior for bayesian face recognition. In *Computer Vision–ECCV 2014*. 2014. 2

[35] J. Lu, G. Wang, W. Deng, and P. Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *Computer Vision–ECCV 2014*. 2014. 2

[36] C. McCool, R. Wallace, M. McLaren, L. El Shafey, and S. Marcel. Session variability modelling for face authentication. *Biometrics, IET*, 2013. 8

[37] D. X. Meina Kan, Shiguang Shan and X. Chen. Side-information based linear discriminant analysis for face recognition. In *British Machine Vision Conference*, 2011. 8

[38] H. Mendez-Vazquez, Y. Martinez-Diaz, and Z. Chai. Volume structured ordinal features with background similarity measure for video face recognition, 2013. 7

[39] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 7

[40] U. Prabhu, J. Heo, and M. Savvides. Unconstrained pose-invariant face recognition using 3d generic elastic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 1, 2

[41] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In *British Machine Vision Conference*, 2013. 6

[42] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep fisher networks for large-scale image classification. *Advances in neural information processing systems*, 2013. 2

[43] V. Struc, J. ganec Gros, S. Dobriek, and N. Pavei. Exploiting representation plurality for robust and efcient face recognition. In *ERK*, 2013. 8

[44] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014. 2

[45] M. Tahir, C. Chan, J. Kittler, and A. Bouridane. Face recognition using multi-scale local phase quantisation and linear regression classifier. In *Image Processing (ICIP), 18th IEEE International Conference on*, 2011. 6

[46] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[47] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1991. 1

[48] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2, 7

[49] L. Wolf and N. Levy. The svm-minus similarity score for video face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 7

[50] J. Wright and G. Hua. Implicit elastic matching with randomized projections for pose-variant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1

[51] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013. 2

[52] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2

[53] C. Zhen, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 7

[54] 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. 1