An Experimental Evaluation of Covariates Effects on Unconstrained Face Verification

Boyu Lu, Jun-Cheng Chen, Member, IEEE, Carlos D. Castillo, Member, IEEE, and Rama Chellappa, Fellow, IEEE

Abstract—Covariates are factors that have a debilitating influence on face verification performance. In this paper, we comprehensively study two covariate related problems for unconstrained face verification: first, how covariates affect the performance of deep neural networks on the large-scale unconstrained face verification problem; second, how to utilize covariates to improve verification performance. To study the first problem, we implement five state-of-the-art deep convolutional networks and evaluate them on three challenging covariates datasets. In total, seven covariates are considered: pose (yaw and roll), age, facial hair, gender, indoor/outdoor, occlusion (nose and mouth visibility, and forehead visibility), and skin tone. We first report the performance of each individual network on the overall protocol and use the score-level fusion method to analyze each covariate. Some of the results confirm and extend the findings of previous studies, and others are new findings that were rarely mentioned previously or did not show consistent trends. For the second problem, we demonstrate that with the assistance of gender information, the quality of a precurated noisy large-scale face dataset for face recognition can be further improved. After retraining the face recognition model using the curated data, performance improvement is observed at low false acceptance rates.

Index Terms—Covariates, deep convolutional neural networks, unconstrained face verification, gender, age, pose.

I. INTRODUCTION

FACE Verification has been receiving consistent attention in computer vision community for over two decades. The task of face verification is to verify whether a given pair of face images/templates belongs to the same subject. Recently, due to the rapid development of deep convolutional neural networks (DCNNs), face verification performance has surpassed human performance in most controlled situations and some unconstrained cases [10], [26], [27], [29], [30], [33], [47], [50]. Although deep features have proven to be more robust to moderate variations in pose, aging, occlusion and other factors than hand-crafted features, some recent works have noticed that face verification performance is still significantly affected by many covariates [28], [40], [43], [48]. Covariates are factors that usually have an undesirable influence on face verification performance (e.g., gender induces different human facial appearance characteristics in nature). Some covariates represent different aspects of faces such as pose, expression and age, some covariates represent subject-specific intrinsic characteristics like gender, race and skin tone, and other covariates reflect extrinsic factors of images, such as illuminations, occlusion and resolution. Analyzing the effects of these covariates can not only help understand fundamental challenges in face verification, but also provide insights for improving existing face verification algorithms.

Previous studies have analyzed many covariates effects on face recognition performance [1], [6], [34], [39]. However, most of them are outdated, and there are several reasons why a new study on these covariates is needed. First, most studies have been conducted before the emergence of deep networks [34]. Since deep networks have significantly improved the robustness of features against many covariates, it is unclear whether the results of covariate effects concluded for hand-crafted features are still relevant when deep features are used. Second, most datasets studied in previous works are small, and the class distributions of some covariates are severely imbalanced [6]. In this situation, some conclusions may become statistically biased. Moreover, due to insufficient data, very few experiments have studied covariate effects at extremely low FARs ($10^{-5}$, $10^{-6}$). Third, face images in former studies were captured in a constrained environment, which is less applicable in practice. Last but not least, most existing papers only focus on whether some covariate values have advantages over other values (e.g., whether a male is easier to recognize than a female), but few of them try to exploit covariate information to improve face verification performance. In fact, some covariates (e.g., gender, race) contain subject-specific information of faces, and are more robust to many extrinsic variations than low-level features. Proper
exploiting them could measurably improve face verification performance [20].

In this paper, we investigate two important problems: a) how different covariates affect the performance of state-of-the-art DCNNs for unconstrained face verification; b) how to utilize covariate information to improve face verification performance. For the first problem, we implement five state-of-the-art face DCNNs and evaluate them on three challenging covariate protocols: 1:1 covariate protocol of the IARPA JANUS Benchmark B (IJB-B) dataset [53] and its extended version the IARPA JANUS Benchmark C (IJB-C) [36], and Celebrity Frontal-Profile Face datasets [46]. We report the performance of each individual network on the overall protocol and use the score-level fusion method to analyze each covariate. We also compare the results with some other well-known and publicly available deep face networks [41], [51]. Among the datasets, IJB-C 1:1 covariate protocol is currently the largest public covariate dataset for unconstrained face verification. Moreover, the IJB-C dataset is designed to have a more uniform geographic distribution of subjects across the globe, which makes it possible to carefully evaluate many covariates (e.g., like age and skin tone) in a comprehensive manner.

By conducting extensive experiments on IJB-B and IJB-C datasets, we observe many interesting behaviors for different covariates. Some of our findings support conclusions drawn from previous studies. For example, extreme yaw angles do substantially degrade the performance [46] and outdoor images are harder to be recognized [24]. Meanwhile, we also find some results which extend the findings of previous works due to the availability of larger datasets. For instance, most previous studies show that face recognition algorithms usually achieve better performance on older subjects than younger subjects [6], [34]. But in their studies, most of the enrolled subjects are under 40 years old. However, our experiments with significantly more subjects with a wider age range show that the performance does not monotonically increase as age progresses. The performance increases from age group [0, 19] to age group [35, 49] but begins to drop for age group [50, 65] and 65+. The results demonstrate that neither too young nor too old people are easy to recognize, but the recognition results for very young people (i.e., [0, 19]) are the worst. Moreover, we are able to better evaluate some covariates like gender where previous works reached contradictory conclusions [34]. Our experiments show that in general, males are easier to match than females. However, when we combine gender with other covariates (age, skin tone) to investigate their mixed effects, we find that the face verification performance for females becomes better than males’ for older age group and darker skin tones. Finally, some of our results are surprising yet rarely analyzed in other papers. One example is that roll variations greatly affect verification performance in unconstrained situation. Since most previous studies may have used manually aligned faces, roll variation was not a significant factor in their studies. However, in unconstrained environments, face alignment becomes a key component and our finding shows that the performance variations might result from the fact that face alignment algorithms fail to work perfectly for faces in extreme roll angles.

For the second problem, we utilize gender information to curate a noisy large-scale face dataset. Specifically, we find that the curated MS-Celeb1M [17], [23] still contains many noisy labels where some subjects still contain images from different genders. Training using the noisy data may potentially hurt the discriminative capability of deep models and degrade their performance, especially in low FAR regions (10^{-5}, 10^{-6}, etc). Therefore, we leverage gender information to further curate the training set and remove subjects mixed with images of both males and females. After retraining the model using the curated data, the performance improves at low FARs. The main contributions of this paper are summarized as follows:

- We comprehensively study the effects of seven covariates on the performance for unconstrained face verification. We test all the covariates using state-of-the-art deep models. This gives insights into the limitations of many existing DCNNs for face covariates.
- We study the mixed effects of multiple covariates. This is an important problem for unconstrained face verification yet not deeply explored by previous studies.
- We propose to utilize gender information to curate the training data and achieve enhanced performance.

The rest of the paper is organized as follows. A brief review of literatures on covariate analysis for face verification is presented in Section II. In Section III, we introduce five state-of-the-art DCNNs for face verification and ways to fuse the similarity scores from them. A method for how to utilize gender covariate for training set curation is presented in Section IV. Experimental results on three different covariates datasets are shown in Section V and finally we summarize and conclude the paper in Section VI.

II. RELATED WORKS

Several prior works discussed the effects of covariates on face recognition performance [1], [6], [7], [13], [14], [34], [37]. Beveridge et al. [6], [7] applied a statistical approach called the Generalized Linear Mixed Model (GLMM) to analyze two types of covariates: subject covariates (e.g., gender, race, wearing glasses) and image covariates (e.g., image size ratio, the number of pixels between eyes). Three algorithms were tested, and they claimed that effects of covariates for different algorithms varied significantly. Givens et al. [13] split faces into three groups (good, bad and ugly) based on the performance of their verification rates. They used GLMM to analyze the underlying effects of different covariates over these three groups. They showed that many covariate effects on verification performance are universal across three groups. Different from the previous works that use statistical methods to analyze the covariates, Lui et al. [34] presented a meta-analysis for six covariates on face recognition performance by summarizing and comparing different papers. In order to guarantee that the conclusions are meaningful, they restricted their analysis to frontal, still, visible light images. Abdurrahim et al. [1] reviewed recent research on demographics related covariates (age, race, and gender). They drew similar conclusions as in [34] for most covariates (e.g., age, gender) while they also pay attention to
interactions among demographics covariates. Grm et al. [14] analyzed the effects of some covariates related to image quality (like blur, occlusion, brightness) and model characteristics (like color information). They used the Labeled Face in the Wild (LFW) [21] dataset to synthesize degraded images and compare the robustness of four widely used DCNNs to each covariate. In the following subsections, we briefly review the main findings of related works for each covariate.

A. Pose

Studies on effects of pose variations on face recognition have been reported in [9], [22], [37], and [47]. Pose variations generally involve yaw, roll and pitch. Normally, roll variations can be eliminated by applying face alignment using similarity or affine transform to warp the face into pre-defined canonical coordinates while yaw and pitch variations are much harder to rectify and thus have a larger impact on the performance than roll in face recognition. Recent studies show that even the best deep-learning based face models are still severely affected by large pose variations [22], [47].

B. Age

The effects of age on face verification performance are usually studied in two ways: aging and age groups. Aging effects are best analyzed in cross-age face verification scenario because it tries to recognize faces from different ages for the same subject. This is a challenging problem because for most subjects their face appearance changes tremendously as they become older [5], [31]. In contrast, age group effects refer to the difficulty in recognizing people from different age groups. This study aims to explore whether a certain age group is harder to recognize than other groups [5], [7], [13], [34].

1) Aging: It has been revealed by almost all studies that age variations impair verification performance. However, the effects may not be significant if the age differences are within several months [16]. Although aging effects become substantial if the acquisition time difference exceeds several years [32], there are still some features preserved on faces that can be utilized for face verification [5]. Best-Rowden et al. applied the mixed-effects models to analyze aging effects using a large mugshot dataset. They showed that the average similarity score of genuine pairs decreases significantly with increasing elapsed time between a gallery and probe. However, they found that on average the genuine pairs can still be recognized at FAR = 0.01%, when the elapsed time is no more than 15 years.

2) Age Groups: The effects of age groups have been discussed by many studies. Interestingly, different from many other covariates where different studies show different results, most studies have come to similar conclusions on age group effects: older subjects are usually easier to recognize than younger subjects [1], [6], [7], [34]. However, most of the experiments were conducted in an environment where age distributions are very imbalanced and the number of samples for young people is much larger than old people. The imbalance increases the difficulty of verification for young people. Ho et al. [19] experimented with each age group evenly distributed. They found that the performance for young ages and old ages did not show statistically significant difference.

C. Gender

Gender is one of the intrinsic characteristics of a human face. Studies on the effects of gender on verification performance have led to different conclusions. Lui et al. [34] summarized covariates research papers from 2001 to 2008. Seven studies found men were easier to recognize [6], [7], while five claimed women were easier [6]–[8], and six reported that gender shows no effects on face recognition performance [6], [7], [11], [12]. More recently, Grother et al. [15] evaluated seven commercial algorithms and five of them were more accurate on males. On the other hand, gender is also shown to have correlations with other covariates like age [34]. Phillips et al. [42] reported that performance difference between males and females decreases as people age.

D. Race and Skin Tone

Race and skin tone are also demographic covariates that represent subject-specific characteristics of people. There were several studies on the effects of races and skin tones on face verification performance, but few of them can be clearly interpreted [1], [34]. This is mainly due to the fact that most datasets are very biased with respect to race distribution. In [34], all the datasets they studied contain more Caucasians than East Asians with a ratio of 3 to 1. Therefore, even if East Asians outperform Caucasians in all the cases in [34], it is still hard to conclude that East Asians are easier for verification. In another paper [15], Grother et al. found that the influences of race on the performance are conflicted for different algorithms. African American are more easily recognized than Caucasians by five out of six algorithms. American Indians and Asian are easily recognized by three algorithms but are more difficult for one algorithm. These results may be simply due to different training processes where algorithms are superior for some races over others. There is also one paper studying the influence of skin tones on face verification [4]. Bar-Haim et al. reported that the effects of skin tone on verification performance are not as important as other unique facial features for certain races.

E. Occlusion

Occlusion could be caused by wearing glasses/sunglasses, masks, scarves or by hairstyle (like bangs). It has been widely investigated that occlusion of key facial parts can substantially degrade the verification performance [6], [7], [14], [37], [48], [52]. However, different algorithms are not sensitive to occlusions to the same degree [7], [14], [48]. There is also one study reporting that consistently wearing glasses may help improve verification performance for faces acquired in outdoors [7].

F. Indoor/Outdoor

The effects of indoor/outdoor are related to some other image covariates like illumination, resolution, and blur. Most studies revealed that indoor performance is generally better than outdoor [7], [8], [24], [25], [34]. Moreover, the indoor/outdoor effect is also found to correlate with other covariates. Beveridge et al. [7], [8] reported that recognition
performance under outdoor environments often favors high resolution images while for low resolution images, indoor environments are preferred. Another finding reported in [7] is that indoor/outdoor taxonomy also affects verification performance for different genders and sometimes may even reverse the trends.

G. Facial Hair

Studies on facial hair effects are limited compared to other common covariates. Earlier studies [11], [12] suggested that performance is better when facial hair exists in at least one of the images. However, the underlying reason for this result is unclear because facial hair is not a unique biometric for recognition and can be changed easily.

III. EVALUATION PIPELINE OVERVIEW

In this section, we briefly introduce the five deep networks that we used to perform unconstrained face verification over covariates. Before feeding a face image into these networks, preprocessing steps including face detection, facial landmark detection and face alignment are performed by using the multi-task CNN framework proposed in [44]. More details about the multi-task CNN are provided in Section III-A5. After feature extraction, we applied triplet probabilistic embedding (TPE) [45] on the deep features to further improve the face verification performance. The TPE learns a projection matrix $W$ by minimizing the negative log-likelihood objective function. More details can be found in [45]. The end-to-end system pipeline is illustrated in Figure 1.

A. Deep Representations for Faces

To capture the different characteristics of faces, we use features extracted from five state-of-the-art deep neural networks. These five networks have different architectures and training sets with their own strengths and weaknesses.

1) Training Set Preparation: To train the deep networks, we use UMD-Faces [2],[3], Megaface [38], and MS-Celeb-1M [17]. In addition, we found that directly using the original MS-Celeb-1M dataset for training does not achieve good performance because the labels are very noisy. Therefore, we used a curated version of MS-Celeb-1M dataset which is done using a clustering method introduced in [23]. The curated dataset contains about 3.7 millions face images from 57,440 identities. After curation, many noisy labels are removed while sufficient amount of face images with different variations are retained.

2) CNN-1: This network employs the ResNet-27 model introduced in [51]. We modify the original model by removing the center loss and replacing the softmax loss with the $L_2$-softmax loss introduced in [43]. In addition, we also add one more fully connected layer with 512-D before $L_2$-softmax layer to reduce the feature dimension and the total number of model parameters. For the input size, we change the original size of $112 \times 96$ to $128 \times 128$ for improved face alignment. To train the model, we use a curated version of the MS-Celeb-1M dataset described in Section III-A1, which contains 3.7 million images from 57,440 subjects.

3) CNN-2: The second network uses the ResNet-101 [18] architecture as the base network. CNN-2 is deeper than CNN-1 and accepts larger inputs of dimensions $224 \times 224$. The basic blocks for CNN-2 use bottleneck structures to reduce the number of model parameters and achieve deeper networks given certain memory constraints. Similar to CNN-1, CNN-2 also replaces the original softmax loss with the $L_2$-softmax loss and adds an extra fully connected layer before the $L_2$-softmax layer. CNN-2 is trained using two different training sets and thus two different models are obtained. One model is called CNN-2_S because a small training set is used (curated MS-Celeb-1M dataset) and the other model is called CNN-2_L because it uses a larger training set (curated MS-Celeb-1M dataset, about 300,000 still images from the UMDFaces dataset [3], and about 1.8 million video frames from the UMD-Faces Video dataset [2]).

4) CNN-3: The Inception-ResNet-v2 [49] model is used as the base network. This model combines the inception architecture with residual connections and scaling layers which...
scale down the residuals for more stable training. We also add a 512-D fully connected layer before the last layer. The training set is the same as for CNN-2.

5) CNN-4: This network is based on the all-in-one CNN architecture [44]. The model is trained in a multi-task learning framework which utilizes the correlations among different tasks to learn a more robust model than learning each task individually. The face detection and facial landmark detection branches share the first six layers and have two separate fully connected layers for each task. The face recognition branch consists of seven convolutional layers followed by three fully connected layers. In this paper, we mainly utilize the face detection, facial landmark detection branches for face alignment, and the face recognition branch to generate face features. We also use the gender classification branch to estimate gender probabilities. The same training set used for CNN-1 and CNN-2. S is used for this network.

B. Face Matching and Score Level Fusion

After we obtain the extracted features from the learned deep networks and the embedding matrix W from TPE [45], the similarity score for each pair \( \{x_i, x_j\} \) is computed by simply using the cosine similarity of the two embedded features:

\[
s_{ij} = \frac{(W_{x_i})^T(W_{x_j})}{\|W_{x_i}\|\|W_{x_j}\|}
\]

(1)

In the last stage of the proposed system, we fuse the scores computed from the five networks as the final similarity score. We observe that the similarity scores may become unreliable when the image quality is poor. Meanwhile, we find the face detection score obtained from the face detection branches of the CNN-4 is a good indication of image quality. Figure 2 shows some hard negative pairs with low detection scores but high similarity scores. We notice that the main reason for the high similarity scores is that these pairs are all very blurred and each pair has similar background. To address this issue, we reweight the similarity scores when the face detection scores of the corresponding pairs are low.

\[
s_i = \begin{cases} 
s_{ij}, & \text{if } ds > \text{thr} \\
\alpha s_{ij}, & \text{otherwise},
\end{cases}
\]

(2)

where \( ds \) is the minimum of the detection scores for the pair of faces, \( \alpha \) is the reweight coefficient.

Then we simply average the reweighted similarity scores from the five networks to get the final results

\[
s = \frac{1}{5} \sum s_i.
\]

(3)

![Fig. 2. Examples of hard negative pairs for low detection confidence but have high similarity scores. ds indicate the detection scores for the images and 5 represents similarity score for each pair.](image)

![Fig. 3. Sample images for IJB-B (first row), IJB-C (second row) and CFP (third row) datasets.](image)

IV. PERFORMANCE IMPROVEMENT BY EXPLOITING GENDER INFORMATION

Although many noisy labels are removed after curating the training set using the clustering method as mentioned in Section III-A1, there still exists many noisy labels which cannot be handled by clustering. Moreover, we observe that some clusters are even mixed with different genders. This motivates us to further curate the training set by exploiting the gender information. First, gender probabilities are estimated using the all-in-one CNN network [44] for all the face images in the pre-curated MS-Celeb-1M dataset in Section III-A1. Since gender estimation may become unreliable when gender probabilities are near 0.5, we only consider faces with gender probability greater than 0.6 (male) or smaller than 0.4 (female). For each subject, if the number of faces from the minority gender is more than 3% of the total number of faces, we eliminate the whole subject. In total, we removed 248,059 faces from 4,160 subjects. It is worth mentioning that we also tried other possible criteria for gender-based curation (e.g., only removing images from minority gender, or use other thresholds instead of 3%) but observed performance drop.

V. EXPERIMENTAL RESULTS

To analyze the covariate effects on unconstrained face verification performance, we evaluate the five deep networks on three challenging face datasets that have face verification covariate protocols: the IARPA JANUS Benchmark B (IJB-B) 1:1 covariates [53], the IARPA JANUS Benchmark C (IJB-C) 1:1 covariates [36] and the Celebrities in Frontal-Profile in the Wild (CFP) [46]. The IJB-B and IJB-C 1:1 covariates both contain seven covariate protocols while the CFP dataset mainly focuses on extreme pose variations. For IJB-B and IJB-C, we first report the performance of each individual network on the overall protocol, and then use the score-level fusion method to analyze each covariate.

A. IJB-B and IJB-C 1:1 Covariate Protocol

The IARPA JANUS Benchmark B (IJB-B) dataset [53] is a moderate-scale unconstrained face dataset with face detection, recognition and clustering protocols. It consists of
1845 subjects with human-labeled ground truth face bounding boxes, eye/nose locations, and covariate meta-data such as occlusion, facial hair, and skin tone for 21,798 still images and 55,026 frames from 7,011 videos. The 1:1 covariate protocol of IJB-B aims to analyze the effects of seven different covariates (i.e., pose (yaw and roll), age, facial hair, gender, indoor/outdoor, occlusion (nose and mouth visibility, forehead visibility), and skin tone.) on face verification performance. The protocol has 20,270,277 pairs of templates (3,867,417 positive and 16,402,860 negative pairs) which enables us to evaluate algorithms at low FAR region of ROC curves (e.g., FAR at $10^{-4}$ and $10^{-6}$). Each template contains only one image or a video frame. The IARPA JANUS Benchmark C (IJB-C) dataset [36] is an extended version of the IJB-B dataset, which consists of 3,531 subjects containing 140,739 images and video frames. The 1:1 covariate protocol has 47,404,001 pair of templates (7,819,362 positive and 39,584,639 negative pairs). Some sample images of the IJB-B and IJB-C datasets are shown in Figure 3.

To understand the effects of different covariates on face verification performance, in addition to the identity label (positive or negative) for each pair of templates, covariate labels are also assigned to each pair. To analyze a certain covariate (like gender), all pairs are split into groups based on the value of covariate labels (female, male). The ROC curves are drawn for each group and the performance difference among different groups reflects the effects of the covariates. When we evaluate the general performance of an algorithm, all the pairs are mixed together without specifying separate covariate labels.

**B. Evaluation on the Overall Protocol**

In the following sections, we first present our experimental results on the overall protocol where covariate labels are not involved and then delve into the details of each covariate result.

1) Results for Five Deep Networks and Score-Level Fusion: To compare the performance of five deep networks, we present the ROC curves for each network and their score-level fusion. For detection score-based fusion, threshold $\text{thr}$ is set to 0.75 and the reweighting coefficient $\alpha$ is set to 0.8. We also did a sensitivity analysis on these two parameters, the details of which are included in the supplementary materials.

Figures 4(a) and 4(b) show the performance for IJB-B and IJB-C 1:1 covariates respectively. From both figures, we observe that CNN-2_S and CNN-3 perform very well at high FARs of the ROC curve, but the performance drops rapidly at low FARs. In contrast, CNN-1, and CNN-4 have smoother curves and perform better at low FARs but worse at high FARs. Meanwhile, CNN-2_L shows very strong performance for all FARs and outperforms the other four networks in middle range of FARs ($FAR = 10^{-4}, 10^{-3}$). Moreover, the fusion results of the five networks outperform all individual models, especially at low FAR of the ROC curve for the IJB-C dataset. This demonstrates the complementary behavior of the different models and fusion can always yield some improvements over individual models. By comparing the ROC curves of IJB-B and IJB-C datasets, we can see similar trends when FARs are larger than $10^{-4}$ but the performance for IJB-B drops faster at low FARs of the ROC curve. In addition, at low FARs, different algorithms perform very differently for IJB-C but similarly for the IJB-B dataset. This indicates that the IJB-B dataset contains more hard negative pairs.

2) Performance Improvement by Gender Based Training Set Curation: To test the effectiveness of the dataset curation method discussed in Section IV, we retrain CNN-1 using the training set curated by exploiting gender information and compare with results obtained before curation. From Table I it can be seen that the performance is improved at low FARs of ROC curves after training set curation on both IJB-B and IJB-C datasets. Since the goal of gender-based curation is to improve the model’s capability to distinguish male and female subjects who looks very similar, performance improvements at low FARs are consistent with this goal because it indicates that the model can deal with hard negative pairs in a better way. On the other hand, we notice that the performance improvements on IJB-C are larger than on IJB-B, which means the gender information is more useful to detect the hard negative pairs in IJB-C than in IJB-B.

3) Comparisons With Other Competitve Methods: We also compare our fusion results with some other state-of-the-art methods and two widely used public models are considered: VGG-Face [41] and Center-Face [51]. We used the pretrained models provided by authors to extract features and followed...
their preprocessing steps on face images. As shown in Table II and Table III, our fusion results outperform both VGG Face and Center-Face by large margins. There are two main reasons for this dramatic performance difference. First, we employ deeper models and various architectures to capture different characteristics of faces and conduct score-level fusion to further boost the performance. Second, the training set we use contains more faces with diverse face variations. In order to investigate the effect of using different training sets, we retrain the Center-Face model using the curated MS-Celeb-1M dataset. As illustrated in Table II and Table III, we can see significant improvements in performance compared to the pretrained model, but the proposed fusion method still outperforms the retrained model significantly.

### C. Evaluation on Pose

To evaluate the effects of pose variations on face verification performance, the protocol provides yaw and roll angles for each face. Since we use the average of the features for original face and its mirrored version as the final face representation, this restricts the range of yaw to [0°, 90°] and roll to [0°, 180°]. Based on the yaw difference between a pair of faces, we divide all pairs into four groups: [0°, 15°], [15°, 30°], [30°, 45°], and [45°, 90°]. Similarly, pairs are also divided into four groups based on roll difference: [0°, 15°], [15°, 30°], [30°, 45°], and [45°, 180°]. Due to space limitations, we did not include the IJB-C plots here because they show similar results as JJB-B.

From Figure 5(a), we observe that the yaw difference between a pair of faces significantly affect face verification performance. The ROC curves decrease monotonically as the yaw difference between two faces increases. Moreover, the performance drops much faster when the yaw difference is larger than 30°. This supports the following two findings: a) deep face representations are robust to moderate yaw changes (less than 30°); b) the state-of-the-art deep networks are still sensitive to large yaw variations (larger than 30°). However, when considering the low FARs regions, we find the performances for different groups become similar. In addition to yaw difference between two faces, another key factor that may influence the performance is the absolute yaw value of faces. In other words, even if the yaw difference between two faces is relatively small (less than 15°), the performance may still be affected when the absolute yaw angles for both faces are large. In order to separate this factor from yaw difference, we further split the group of yaw difference [0°, 15°] into four subgroups based on their absolute yaw angles: [0°, 15°], [15°, 30°], [30°, 45°], and [45°, 90°], where the degrees are computed by averaging the absolute yaw angles of a pair of faces. The ROC curves are shown in Figure 5(b). Similar to the effect of yaw difference, the absolute yaw angles of faces larger than 30° cause a large performance drop while performance is not affected much when yaw angles are less than 30°. By comparing Figures 5(a) and 5(b), we have another interesting finding: performance for absolute yaw angles in [45°, 90°] and for yaw difference in [45°, 90°] are comparable, which means that as long as at least one of the two faces is in extreme yaw angle, the performance will be poor. This result demonstrates that face images with extreme yaw angles ([45°, 90°]) are hard for face matching regardless of the yaw difference because a large part of facial information is missing.

Figure 6 shows the face verification performance for various roll difference between two faces. We find that performance is better for groups whose roll differences are smaller than 30°. This result is surprising because in general the roll difference should not affect the face verification performance since 2D

### TABLE I

**Performance Comparison Between Before and After Gender-Based Training Set Curation on IJB-B and IJB-C 1:1 Covariate Protocol. All the Results Are Generated Using the CNN-1 Architecture**

| Method          | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| IJB-B before curation | 0.0252 | 0.1602 | 0.4453 | 0.6282 | 0.7474 | 0.8493 | 0.9328 |
| IJB-B after curation | 0.0249 | 0.1131 | 0.4586 | 0.6214 | 0.7681 | 0.8477 | 0.9250 |
| IJB-C before curation | 0.0167 | 0.1336 | 0.5029 | 0.6493 | 0.7664 | 0.8624 | 0.9364 |
| IJB-C after curation | 0.0261 | 0.3244 | 0.5378 | 0.6686 | 0.7684 | 0.8566 | 0.9337 |

### TABLE II

**Performance Comparison for Different Methods on the IJB-B 1:1 Covariate Overall Protocols. Our Fusion Results Are Generated by Detection Score-Based Fusion of the Five Deep Models. VGG-Face and Center-Face Results Are Derived by Applying Their Pretrained Models to Extract Features and Following the IJB-B 1:1 Covariate Overall Protocol. Center-Face (Retrain) is Retrained Using the Curated MS-Celeb-1M Dataset and the Center-Face Model.**

| Method          | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| VGG-Face        | 0.0150 | 0.0440 | 0.0994 | 0.1518 | 0.2180 | 0.3518 | 0.5723 |
| Center-Face     | 0.0065 | 0.0333 | 0.0780 | 0.1363 | 0.2370 | 0.4206 | 0.7501 |
| Center-Face (retrain) | 0.0917 | 0.1556 | 0.3880 | 0.6014 | 0.7620 | 0.8592 | 0.9460 |
| Fusion of our five model | 0.0396 | 0.1707 | 0.4642 | 0.7093 | 0.8454 | 0.9213 | 0.9688 |

### TABLE III

**Performance Comparison for Different Methods on the IJB-C 1:1 Covariate Overall Protocols. Our Fusion Results Are Generated by Detection Score-Based Fusion of the Five Deep Models. VGG-Face and Center-Face Results Are Derived by Applying Their Pretrained Model to Extract Features and Following the IJB-C 1:1 Covariate Overall Protocol. Center-Face (Retrain) Is Retrained Using the Curated MS-Celeb-1M Dataset and the Center-Face Model.**

| Method          | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ | TAR@FAR = 10⁻¹ | TAR@FAR = 10⁻⁶ |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| VGG-Face        | 0.0813 | 0.0792 | 0.1159 | 0.1818 | 0.2275 | 0.3936 | 0.5918 |
| Center-Face     | 0.0479 | 0.0552 | 0.0905 | 0.1582 | 0.2747 | 0.4397 | 0.7568 |
| Center-Face (retrain) | 0.2417 | 0.3599 | 0.5023 | 0.6401 | 0.7640 | 0.8765 | 0.9441 |
| Fusion of our five model | 0.2571 | 0.5229 | 0.6478 | 0.7623 | 0.8599 | 0.9261 | 0.9761 |
Fig. 5. ROC curves (a) when the yaw difference between two face images changes and (b) when absolute yaw angle of faces changes. The range is from $0^\circ$ to $90^\circ$ because we average the features for original face and its mirrored image as the final face representation. The absolute yaw angles are computed by averaging two faces. The dashed line represents the results for the overall protocol.

Fig. 6. ROC curves when the roll angle difference between two face images changes for IJB-B. The range is from $0^\circ$ to $180^\circ$. The dashed line represents the results for the overall protocol.

face alignment is performed before face matching to normalize all faces to have the same roll angle. However, the performance drop when increasing the roll difference shows that facial landmarks may not be accurate so that faces are not normalized as expected when the roll angle is large.

D. Evaluation on Gender

From Figure 7(a), it can be observed that the performance for men is much better than women on the IJB-B dataset. The results for the IJB-C dataset show similar trends and are not included due to space limitations. A possible explanation for this result is that women’s faces are often occluded by their long hair and their face appearance are changed by makeup. To further investigate the underlying reasons of our observation, we use t-SNE plots [35] to analyze the feature distributions under different genders and the results are illustrated in Figure 8. The small clusters represent different subjects and we also include the t-SNE visualization based on identities in the supplementary materials for reference. We can see that the feature distributions for men are much more separated and discriminative than women, which lead to better performance.

E. Evaluation on Age

The 1:1 covariate protocol labels the test pairs into six categories based on their age distributions. Ages that are different for two faces in a pair are labeled as $-1$. Results for IJB-B dataset are shown in Figure 7(b). Due to space limitations, we do not include the IJB-C plots here because they show similar results as for IJB-B. The dashed line represents performance for the overall protocol while the solid lines present curves for different age groups. It is shown that performance goes up when age increases from 0 to 49. In contrast, the curves begin to fall when age is older than 49. It means the middle-age group ([35, 49]) is the easiest one to be recognized while too young or too old subjects are both challenging for face verification. One possible explanation for this result may be because new born babies all look very similar and their unique facial features begin to emerge as they grow. However, as people age, some common features for elderly people like wrinkles and sagging skins impair the uniqueness of their facial characteristics, which may make them harder to be distinguished. On the other hand, we find the performances for age groups that are older than 35 become closer at low FARs. In addition, we notice that age group $-1$ (ages of two images are different) performs similarly as the overall protocol, which means cross-age face verification is as hard as the general case. Nonetheless, this dataset does not fully explore the difficulty of cross-age face verification because the IJB-B and IJB-C datasets do not have images from the same person across large age gaps.

F. Evaluation on Skin Tone

For skin tone, the protocol defines six classes: (1) light pink, (2) light yellow, (3) medium pink/brown, (4) medium yellow/brown, (5) medium dark brown, and (6) dark brown. From Figure 9, we observe that the performances for different skin tone groups show different trends on IJB-B and IJB-C. For IJB-B, the ROC curves for different groups are well separated. A general trend is that performance falls when the skin tone becomes darker. However, a counterexample is skin tone group 6 (darkest), which performs better than group 2 to group 5. On the other hand, the performance for group 3 drops rapidly and performs the worst at low FARs. This demonstrates that the hard negative pairs for group 3 are more difficult to recognize. For IJB-C, except group 1 and group 5 which have
the same trends as IJB-B, the performances for other skin tone groups are very close. Thus, we can only draw the conclusion that skin tone group 1 is the easiest and skin tone group 5 is the hardest for face verification. However, since defining or recognizing skin tones is ambiguous sometimes, it is hard to decide which skin tone is easier for face verification only from these results. In Figure 12, we visualize the feature distribution for different skin-tone groups in the IJB-B dataset. We can easily find that features for group 1 (shown in red dots) are most separated and thus achieve the best performance. Nonetheless, feature distributions for other groups do not show much information.

G. Evaluation on Mouth and Nose, and Forehead Visibility

To evaluate the effects of occlusion, the protocol tests two types of visibilities: mouth and nose visibility, and forehead visibility. Label 0 (1) represents the parts are both invisible (visible) for two images, and label $-1$ means the part is visible for one image but not for the other. The ROC curves for IJB-B dataset are presented in Figures 10(a) and 10(b) respectively. We see similar results for mouth/nose and forehead visibility: class $-1$ and 0 have comparable performance but are worse than class 1, which means that performance falls by large margins if nose, mouth or forehead are occluded for at least one of the images. This result indicates the importance of the visibility of key facial parts for recognizing faces. However, when considering the low FARs regions, we find the performances for different groups become similar. This means for low FAR regions, occlusion is not the key factor that decide performance since the pairs are often affected by many covariates (e.g., pose, occlusion, illumination).

H. Evaluation on Facial Hair

There are four classes for evaluation in facial hair protocol: no facial hair, moustache, goatee and beard respectively. Label $-1$ means facial hair classes are different for two images. From Figure 11(a), we observe that performance is not very sensitive to facial hair changes. This result demonstrates that facial hair does not change the key features of faces and state-of-the-art deep models can handle most facial hair variations.

I. Evaluation on Indoor/Outdoor

The last covariate we evaluate in the protocol is indoor/outdoor. Outdoor is labeled as 0 and indoor is 1. Label $-1$ means one image is taken indoor and the other outdoor. Performance is shown in Figure 11(b). We can see that the performance of class 1 is much better than class 0 and $-1$. This implies that indoor images are easier for face verification. Different from occlusion, we find that performance for indoor is still better than outdoor even at low FARs. This leads to a claim that indoor is an important condition to recognize hard negative pairs. There are two possible reasons for this result. First, outdoor images could be easily over-exposed and lose significant facial information. Second, outdoor images are often taken by hand-held cameras when people are walking. In contrast, indoor images are usually captured without much motion. So the image quality for indoor images is often better than outdoor images.

J. Evaluation on the Effects of Multiple Covariates

In unconstrained face verification, multiple face covariates are often correlated with each other which may affect the performance. It has been found that some covariates may show
different trends on face verification performance when other covariates are considered together [7], [42]. To study the correlations among the different covariates, we chose four pairs of related covariates and evaluated their interactive effects: gender and age, gender and skin tone, indoor (outdoor) and nose-mouth visibility, indoor (outdoor) and yaw angle difference. Due to space limitation, all experimental results are reported for the IJB-B dataset.

1) Evaluation on Gender and Age: In order to show how gender and age influence each other, we draw ROC curves in Figure 13(a) for each possible combination of values from genders and age groups. Different age groups are represented using different colors and men/women is showed in solid/dashed lines. First, we fix the gender factor and compare the performance of different age groups for males or females. We see that males and females show very different
trends on age group effects. More specifically, men in middle age group [35, 49] performs best and the performances for men in age group [50, 64] and 65+ decrease. In contrast, for women the performance always increases when age groups get older.

Alternatively, we can fix the age group factor and compare the performance of men and women for each age group. As observed in Section V-D, in general, results for men are better than those for women. However, this finding does not hold for age group [50, 64] and 65+. For age group [50, 64], men and women perform comparably while for age group 65+ women outperform men.

2) Evaluation on Gender and Skin Tone: We repeated the procedure discussed above for analyzing the combination of gender and skin tone. The ROC curves are shown in Figure 13(b). For skin tone groups 4 and 6, performance for women is better than that for men, while men perform better for group 1, 2 and 5. For skin tone group 3, men and women perform similarly. This result shows that the combinations of gender and skin tone do not show clear trends and the performance is dependent on datasets.

3) Evaluation on Indoor (Outdoor) and Nose-Mouth Visibility: In addition to the demographic covariates, we are also interested in the mixed effects of covariates related to extrinsic factors. Figure 14(a) shows the performance for different indoor/outdoor and nose-mouth visibility combinations. As we already saw, visible nose-mouth and indoor are more favorable for better performance. However, these two factors may not have independent impacts on performance. From Figure 14(a), we find that only when nose or mouth is visible and the images are taken indoor, the performance is good. Either occlusion or outdoor can deteriorate the performance. At low FARs, we find that indoor/outdoor is more important than nose-mouth visibility, as the performance for green dashed line is better than yellow solid line in this region. This finding confirms the claim in Section V-G and V-I.

4) Evaluation on Indoor (Outdoor) and Yaw Angle Difference: The last combination we considered is indoor/outdoor and yaw angle difference. The ROC curves are presented in Figure 14(b). We notice that when fixing the indoor/outdoor factor, the performance for smaller yaw angle difference is always better. On the other hand, when the yaw angle difference is fixed, indoor faces always outperform outdoor faces. This result demonstrates that yaw angle difference and indoor/outdoor can affect the face verification performance independently and changing any one of the two factors can affect the performance.

K. Evaluation on CFP Dataset

Since pose variation is a key challenging issue for face verification, we also used the Celebrities in Frontal-Profile (CFP) dataset to further investigate the underlying effects of extreme pose variations on unconstrained face verification performance. The CFP dataset consists of 7,000 still images from 500 subjects with 14 images per subject. For each subject, it has 10 images in frontal pose and 4 images in profile pose. To evaluate the performance for different poses, the protocol contains two settings: frontal-to-frontal (FF) and frontal-to-profile (FP) face verification. In the frontal-to-frontal setting, two test images are both in frontal pose and in frontal-to-profile setting, a test pair includes one frontal face and one profile face. Each setting divides the whole dataset into ten splits and each split consists of 350 positive and 350 negative pairs. Some sample images are shown in Figure 3.

1) Performance Evaluation Metrics: We follow the performance evaluation metrics used in [46] and report three numbers for each setting: Area under the curve (AUC), Equal Error Rate (EER) and Accuracy. AUC measures the area under ROC curves and ranges from 0 to 1 where higher value corresponds to better performance. EER is the point where the false accept rate is equal to false reject rate. It ranges from 0 to 1 with lower values indicating better performance. We use an optimal threshold to classify all pairs and calculate the classification accuracy. For the optimal threshold, we chose the value that provides highest classification accuracy on the cross validation set.

2) Results for Frontal-to-Frontal and Frontal-to-Profile Protocols: The experimental results for frontal-to-frontal and frontal-to-profile protocols are summarized in Table IV. CNN-1 to CNN-4 results are obtained by using the same models and same processing steps for IJB-B and IJB-C experiments. For the fusion part, since all detection scores for the images in CFP dataset is near 1, we simply average the similarity score for CNN-1 through CNN-4. Deep features and human results are directly cited from [46]. The performance is reported by averaging over ten splits.

For the frontal-to-frontal setting, CNN-1 to CNN-4 all outperform both the deep features method and human performance in [46]. CNN-2_S and CNN-2_L perform almost identically. CNN-2 and CNN-3 perform similarly and their performances are slightly better than CNN-1 and CNN-4. Since performances of CNN-2 and CNN-3 have already saturated, fusion results for the five networks do not change much compared to CNN-2 or CNN-3. For the frontal-to-profile setting, different algorithms begin to show significant difference in performance. CNN-1 results are slightly worse than human performance but are 2% better than CNN-4. On the other hand, CNN-2 and CNN-3 both surpass human performance by more than 2%. Another interesting finding is that while
Fig. 13. ROC curves corresponding to age and gender (left) changes, and skin tone and gender (right) changes. Color lines represent different age groups and skin tones where small numbers represent light skin tones. Results for women are showed in dashed lines and solid lines represent results for men.

Fig. 14. ROC curves corresponding to nose-mouth visibility and indoor/outdoor (left), and yaw difference and indoor/outdoor. Outdoor is shown in dashed lines and solid lines represent indoor.

| TABLE IV  |
|------------------|------------------|------------------|------------------|------------------|
|                | Frontal-to-Frontal | Frontal-to-Proile |                |
|                | Accuracy | EER    | AUC    | Accuracy | EER    | AUC    |
| Deep features  |          |        |        |          |        |        |
| Human          | 0.964(0.001) | 0.035(0.007) | 0.994(0.003) | 0.849(0.011) | 0.150(0.020) | 0.800(0.016) |
| CNN-1          | 0.988(0.002) | 0.012(0.004) | 0.999(0.001) | 0.946(0.011) | 0.050(0.011) | 0.989(0.005) |
| CNN-2_S        | 0.997(0.003) | 0.003(0.003) | 1.000(0.000) | 0.981(0.007) | 0.018(0.007) | 0.997(0.002) |
| CNN-2_F        | 0.996(0.003) | 0.004(0.003) | 1.000(0.000) | 0.980(0.004) | 0.021(0.006) | 0.997(0.002) |
| CNN-3          | 0.994(0.004) | 0.006(0.005) | 1.000(0.001) | 0.969(0.009) | 0.029(0.011) | 0.994(0.003) |
| CNN-4          | 0.982(0.008) | 0.018(0.008) | 0.998(0.001) | 0.912(0.012) | 0.085(0.012) | 0.972(0.006) |
| Fusion         | 0.995(0.003) | 0.004(0.004) | 1.000(0.001) | 0.973(0.008) | 0.027(0.008) | 0.996(0.002) |

the performance for different algorithms do not vary much in frontal-to-frontal protocol, the performance drops from frontal-to-frontal to frontal-to-profile is quite different among the compared algorithms. Generally speaking, better algorithms are more robust to extreme yaw variations and always have smaller performance degradation for frontal-to-profile setting. In particular, CNN-2_S has the smallest performance drop of 1.6% from frontal-to-frontal to frontal-to-profile, which is similar to human performance. However, if we compare the results with Section V-C, even the best results are still severely affected by pose variations. This is because the IJB-B and IJB-C datasets contain other challenging factors and pose variations can still degrade performance once combined with these factors. Therefore, even for state-of-the-art face models, there is still room to improve robustness to extreme pose variations.

VI. CONCLUSION AND FUTURE WORK

In this paper, we report the results of comprehensive experiments performed to study the effects of covariates on unconstrained face verification performance. Our evaluations are based on deep learning networks and large training data sets. We also curate the training data by exploiting gender information and achieve improved performance. Experimental results on the overall protocols of IJB-B and IJB-C covariate verification tasks show the outstanding performance of five implemented deep models and their score-level fusion.
However, when we focus on each specific covariate, we find that many covariates still significantly affect the verification performance. Pose variations and occlusions are the top confounding factors that could cause performance drop by large margins. Indoor performance is much better than outdoors. On the other hand, the difficulty of unconstrained face verification varies significantly for different demographic groups. Age, gender and skin tone all have shown impacts on performance. Specifically, males are easier to verify than females and old subjects generally performs better than young ones. For skin tone, light pink achieves the best performance while medium-dark brown performs the worst. However, since IJB-B and IJB-C show very different tendencies on skin tone groups, we are not be able to draw a clear conclusion on its effects.

Most of the findings discussed above confirm the conclusions of previous studies. However, there are also some new findings that were rarely mentioned by other studies or somewhat surprising. First, we find that verification performance does not increase monotonically as subjects get older. In contrast, performance begins to drop for age group of [50, 65] and 65+. This result is different from most studies which claim older subjects are always easier to be recognized. However, since most of other studies did not have a sufficient number of older subjects to analyze, their results still make sense because middle age group performs better than children and teenagers. Second, we observed that extreme roll angle differences between faces still affect performance substantially. This result is unexpected as roll variations should be eliminated by face alignment. Therefore, we conclude that face alignment performance needs to get better when faces are in extreme roll angles.

Finally, we investigated the mixed effects of multiple covariates. First, males and females show very different trends on the effects of age groups. For males, performance first increases then drops when age goes up while for females, older age groups always perform better. On the other hand, the interaction between gender and skin tone does not show clear trends. Second, when we consider indoor/outdoor and occlusion together, we find that indoor and nose-mouth visibility must be satisfied simultaneously to achieve good performance. However, indoor/outdoor and yaw angle difference can affect the performance independently.

Some of the results from our studies show several promising research directions. First, apart from the yaw problem, we should also consider the influence of roll when designing face verification systems. This can be done by either improved face alignment or more robust feature extraction models. Second, since gender, age and skin tone all have significant impact on performance, we may collect the training set more carefully to improve the performance on certain demographic groups. Third, we show preliminary results on how to use gender estimation for training data curation. Other covariates like race may also be used in a similar way.

REFERENCES

[1] S. H. Abdurahim, S. A. Samad, and A. B. Huddin, “Review on the effects of age, gender, and race demographics on automatic face recognition,” Vis. Comput., vol. 34, no. 11, pp. 1617–1630, 2018.
[2] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa, “The Do’s and Don’ts for CNN-based face verification,” arXiv preprint arXiv:1705.07426, 2017.
[3] A. Bansal, A. Nanduri, C. Castillo, R. Ranjan, and R. Chellappa, “UMDFaces: An annotated face dataset for training deep networks,” arXiv preprint arXiv:1611.01484, 2016.
[4] Y. Bar-Haim, T. Saidel, and G. Yovel, “The role of skin colour in face recognition,” Perception, vol. 38, no. 1, pp. 145–148, 2009.
[5] L. Best-Rowden and A. K. Jain, “Longitudinal study of automatic face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 1, pp. 148–162, Jan. 2018.
[6] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper, “Factors that influence algorithm performance in the face recognition grand challenge,” Comput. Vis. Image Understand., vol. 113, no. 6, pp. 750–762, 2009.
[7] J. R. Beveridge et al., “FRVT 2006: Quo vadis face quality?,” Image Vis. Comput., vol. 28, no. 5, pp. 732–743, 2010.
[8] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui, “Focus on quality, predicting FBVT 2006 performance,” in Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), 2008, pp. 1–8.
[9] C. Ding and D. Tao, “A comprehensive survey on pose-invariant face recognition,” ACM Trans. Intell. Syst. Technol., vol. 7, no. 3, p. 37, 2016.
[10] Y. Duan, J. Lu, J. Feng, and J. Zhou, “Context-aware local binary feature learning for face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1139–1155, May 2018.
[11] G. Givens, J. R. Beveridge, B. A. Draper, and D. Bolme, “A statistical assessment of subject factors in the PCA recognition of human faces,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW), vol. 8, 2003, p. 96.
[12] G. Givens, J. R. Beveridge, B. A. Draper, P. Grother, and J. P. Phillips, “How features of the human face affect recognition: A statistical comparison of three face recognition algorithms,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 2, 2004, p. 2.
[13] G. H. Givens et al., “Introduction to face recognition and evaluation of algorithm performance,” Comput. Stat. Data Anal., vol. 67, pp. 236–247, Nov. 2013.
[14] K. Grim, V. Struc, A. Artiges, M. Caron, and H. K. Ekenel, “Strengths and weaknesses of deep learning models for face recognition against image degradations,” IET Biometrics, vol. 7, no. 1, pp. 81–89, 2017.
[15] P. J. Grother, G. W. Quinn, and P. J. Phillips, “Report on the evaluation of 2D still-image face recognition algorithms,” NIST, Gaithersburg, MD, USA, Rep. 7709, p. 106, 2010.
[16] G. Guo, G. Mu, and K. Ratanek, “Cross-age face recognition on a very large database: The performance versus age intervals and improvement using soft biometric traits,” in Proc. 20th Int. Conf. Pattern Recognit. (ICPR), 2010, pp. 3392–3395.
[17] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “MS-Celeb-1M: A dataset and benchmark for large scale face recognition,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 1–16.
[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[19] W. H. Ho, P. Watters, and D. Verity, “Are younger people more difficult to identify or just a peer-to-peer effect,” in Proc. Int. Conf. Comput. Anal. Images Patterns, 2007, pp. 351–359.
[20] G. Hu et al., “Attribute-enhanced face recognition with neural tensor fusion networks,” in Proc. ICCV, 2017, pp. 3764–3773.
[21] 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,” Comput. Sci. Dept., Univ. Massachusetts, Amherst, MA, USA, Rep. 07-49, Oct. 2007.
[22] X. Li et al., “Maximally correlated correspondence estimation for face recognition across pose,” IEEE Trans. Image Process., vol. 23, no. 10, pp. 4587–4600, Oct. 2014.
[23] W.-A. Lin, J.-C. Chen, and R. Chellappa, “A proximity-aware hierarchical clustering of faces,” arXiv preprint arXiv:1703.04835, 2017.
[24] Z. Liu and S. Sarkar, “Outdoor recognition at a distance by fusing gait and face,” Image Vis. Comput., vol. 25, no. 6, pp. 817–832, 2007.
[25] B. Lu, R. Chellappa, and N. M. Nasrabadi, “Incremental dictionary learning for unsupervised domain adaptation,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2015, pp. 1–12.
[26] B. Lu, J.-C. Chen, and R. Chellappa, “Regularized metric adaptation for unconstrained face verification,” in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), 2016, pp. 4112–4117.
[27] B. Lu and N. M. Nasrabadi, “Task-driven dictionary learning with different Laplacian priors for hyperspectral image classification,” in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), 2015, pp. 2868–2871.
[28] B. Lu, J. Zheng, J.-C. Chen, and R. Chellappa, “Pose-robust face verification by exploiting competing tasks,” in Proc. IEEE Winn. Conf. Appl. Comput. Vis. (WACV), 2017, pp. 1124–1132.

[29] J. Lu, J. Hu, and Y.-P. Tan, “Discriminative deep metric learning for face and kinship verification,” IEEE Trans. Image Process., vol. 26, no. 9, pp. 4269–4282, Sep. 2017.

[30] J. Lu, J. Hu, and J. Zhou, “Deep metric learning for visual understanding: An overview of recent advances,” IEEE Signal Process. Mag., vol. 34, no. 6, pp. 76–84, Nov. 2017.

[31] J. Lu, V. E. Liong, and J. Zhou, “Cost-sensitive local binary feature learning for facial age estimation,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5356–5368, Dec. 2015.

[32] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, “Learning compact binary face descriptor for face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 10, pp. 2041–2056, Oct. 2015.

[33] J. Lu, G. Wang, and J. Zhou, “Simultaneous feature and dictionary learning for image set based face recognition,” IEEE Trans. Image Process., vol. 26, no. 8, pp. 4042–4054, Aug. 2017.

[34] Y. M. Lui et al., “A meta-analysis of face recognition covariates,” in Proc. IEEE 3rd Int. Conf. Biometrics Theory Appl. Syst. (BTAS), 2009, pp. 1–8.

[35] L. V. D. Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, Nov. 2008.

[36] B. Maze et al., “IARPA janus benchmark—C: Face dataset and protocol,” in Proc. IAPR Int. Conf. Biometrics, 2018, pp. 158–165.

[37] M. M. Ghazi and H. K. Ekenel, “A comprehensive analysis of deep learning based representation for face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2016, pp. 34–41.

[38] A. Nech and I. Kemelmacher-Shlizerman, “Level playing field for million scale face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3406–3415.

[39] J. Paone, S. Biswas, G. Aggarwal, and P. Flynn, “Difficult imaging covariates or difficult subjects? An empirical investigation,” in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), 2011, pp. 1–8.

[40] C. J. Parde et al., “Face and image representation in deep CNN features,” in Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), Washington, DC, USA, 2017, pp. 673–680.

[41] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in Proc. Brit. Mach. Vis. Conf., 2015, pp. 41.1–41.12.

[42] P. J. Phillips et al., “Evaluation report,” 2003.

[43] R. Ranjan, C. D. Castillo, and R. Chellappa, “L2-constrained softmax loss for discriminative face verification,” arXiv preprint arXiv:1703.09507, 2017.

[44] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, “An all-in-one convolutional neural network for face analysis,” in Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), 2017, pp. 17–24.

[45] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, “Triplet probabilistic embedding for face verification and clustering,” in Proc. IEEE 8th Int. Conf. Biometrics Theory Appl. Syst. (BTAS), 2016, pp. 1–8.

[46] S. Sengupta et al., “Frontal to profile face verification in the wild,” in Proc. IEEE Winn. Conf. Appl. Comput. Vis. (WACV), 2016, pp. 1–9.

[47] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Columbus, OH, USA, 2014, pp. 1891–1898.

[48] Y. Sun, X. Wang, and X. Tang, “Deeply learned face representations are sparse, selective, and robust,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, 2015, pp. 2892–2900.

[49] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in Proc. AAAI, 2017, pp. 4274–4284.

[50] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1701–1708.

[51] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 499–515.

[52] R. Weng, J. Lu, and Y.-P. Tan, “Robust point set matching for partial face recognition,” IEEE Trans. Image Process., vol. 25, no. 3, pp. 1163–1176, Mar. 2016.

[53] C. Whitelam et al., “IARPA janus benchmark-B face dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, Honolulu, HI, USA, Jul. 2017, pp. 592–600.

Boyu Lu is currently a Research Assistant with the University of Maryland, College Park. His research interests include face recognition, deep learning, and domain adaptation.

Jun-Cheng Chen received the Ph.D. degree from the University of Maryland, College Park, in 2016. He is a Post-Doctoral Research Fellow with the University of Maryland Institute for Advanced Computer Studies. His current research interests include computer vision and machine learning with applications to face recognition and facial analysis. He was a recipient of the ACM Multimedia Best Technical Full Paper Award in 2006.

Carlos D. Castillo received the Ph.D. degree from the University of Maryland, College Park, in 2012. He is an Assistant Research Scientist with the University of Maryland Institute for Advanced Computer Studies. His current research interests include stereo matching, multiview geometry, face detection, and alignment and recognition.

Rama Chellappa is a Distinguished University Professor, a Minta Martin Professor of engineering, and the Chair of the ECE Department with the University of Maryland, where he received college and university level recognitions for research, teaching, innovation, and mentoring of undergraduate students. He holds four patents. He was a recipient of the K. S. Fu Prize from the International Association of Pattern Recognition, the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society, four IBM Faculty Development Awards, and the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. In 2010, he was recognized as an Outstanding ECE by Purdue University. He served as the Editor-in-Chief of PAMI. He is a Golden Core Member of the IEEE Computer Society, served as a Distinguished Lecturer of the IEEE Signal Processing Society and as the President of IEEE Biometrics Council. He is a fellow of IAPR, OSA, AAAS, ACM, and AAAI.