Occlusion-guided compact template learning for ensemble deep network-based pose-invariant face recognition

Yuhang Wu and Ioannis A. Kakadiaris
Computational Biomedicine Lab, University of Houston
{yu35,ikakadia}@central.uh.edu

Abstract

Concatenation of the deep network representations extracted from different facial patches helps to improve face recognition performance. However, the concatenated facial template increases in size and contains redundant information. Previous solutions aim to reduce the dimension of the facial template without considering the occlusion pattern of the facial patches. In this paper, we propose an occlusion-guided compact template learning (OGCTL) approach that only uses the information from visible patches to construct the compact template. The compact face representation is not sensitive to the number of patches that are used to construct the facial template, and more suitable for incorporating the information from different view angles for image-set based face recognition. Different from previous ensemble models that use occlusion masks in face matching (e.g., DPRFS [38]), the proposed method uses occlusion masks in template construction and achieves significantly better image-set based face verification performance on challenging database with a template size that is an order-of-magnitude smaller than DPRFS.

1. Introduction

The problem of template learning in face recognition is to map a single facial image into a vector representation so that the distance between these representations reflects the actual similarity between facial images (or videos). In recent work, researchers have focused more on improving the discriminative capability of the template, and paid relatively less attention to the size of the template. The size of the template becomes important for large-scale face retrieval when the gallery database contains millions of images [34]. A real-world surveillance scenario becomes infeasible to compute the similarity scores between one probe and all images in the gallery when it takes minutes to process a single video frame. Another application scenario is biometric credit cards, where the size of the template is limited by the memory of the chip inside it. In this work, we focus on learning a compact floating point template whose vector length should be as small as possible while preserving its discriminative capability.

Learning a compact but highly discriminative template is challenging; the recent research Gong et al. [9] indicates that template length directly impacts the discriminative capability of the template. To gather as much discriminative information as possible in the template, our starting point is based on an ensemble deep neural network model named DPRFS presented by Xiang et al. [38]. Like other ensemble deep neural network models [32, 30, 31, 13, 23, 4], DPRFS concatenates facial representations from multiple facial patches, but uses an addition occlusion vector to indicate which facial patch was occluded. It was demonstrated that the template generated by DPRFS achieves state-of-the-art results for single image based face matching under large head pose variation [38], but as the output of an ensemble model, it has the following limitations: (i) the template size is too big (16K bytes) for face retrieval and stor-
age, (ii) face matching speed is limited by selecting non-occluded face representations for similarity computations, (iii) it is hard to generate a single template that takes into account all the occlusion patterns of images in an image set.

To address the aforementioned limitations of the ensemble model, we propose an occlusion-guided compact template learning method to encourage the information integration between the face representations that are generated by each individual deep network. This operation preserves the discriminative co-occurrence pattern between multiple facial regions and significantly improves the performance of a compact template. The main contributions of this paper are: (i) we propose an occlusion-guided neural network architecture to learn a compact facial template from the face representations and occlusion masks generated by an ensemble model, (ii) the compact template can be generated from different number of visible facial patches.

2. Related work

In the literature on facial template learning, researchers have paid more attention to the performance of the templates rather than the template size. To the best of our knowledge, the size of the smallest template that achieves more than 99.5% verification rate on LFW database [14] is 0.5 KB [23, 29] (The data type of the template is assumed to be 32-bit single precision floating point as used in standard convolutional neural networks (CNN) [10]). Although most of the selected algorithms reported near saturated verification performance on LFW database, their performances degrade in the databases that contain large head pose variations (e.g., UHDB31 database [21]). Figure 2 indicates the performance of templates generated by multiple methods (COTS v1.9, FaceNet [29] (open-sourced version), VGG-face [26], and DPRFS [38]) on UHDB31 database. It also highlighted the performance of the template generated by the proposed algorithm OGCTL.

Many of the recent face recognition systems employ ensemble models [32, 30, 31, 13, 23, 4, 2], where face representations of multiple convolutional neural networks are concatenated together as the final facial template to enrich the template’s representation capability. Hu et al. [13] used extensive experiments to show that the concatenated representation consistently outperforms the output of a stand-alone CNN model. To learn diverse facial representation for concatenation, a common solution is to divide the facial image into multiple patches [32, 30, 31, 13, 23, 4], and for each patch learn a separate CNN. However, these methods do not take into account the occlusion of patches under large headpose variations. Some of the patches should be de-activated on-line due to facial occlusion because their corresponding features in the facial template are no longer informative. Using these non-discriminative face representations in face

Figure 2: Rank-1 identification rate vs. template size on UHDB31 database.

matching is detrimental to the final face recognition performance.

A compact template can be further compressed by the principal component analysis (PCA) [17], product quantization [16], or hashing-based binarization approaches [5, 22]. PCA is a well-known dimension reduction approach which has been applied in [32, 30] to reduce facial template size from 76.8 KB to 0.625 KB. Wang [24] employs Product Quantization [16] to convert a 1.25 KB into a binary 64 Bytes template for large-scale face retrieval. Recently, deep Hashing-based approaches have the capability to generate a binary template from the output of CNN with hashing layers [5, 22]. The common point of all these approaches is that they require a floating point face representations as the starting point to derive a more compact facial template. The contribution of our approach is to provide a method to generate a compact face representation (0.5 KB to 1 KB) as a good candidate for aforementioned studies [17, 16, 5, 22], and this candidate is robust to facial self-occlusion caused by large head pose variations.

3. Method

3.1. Template and occlusion mask

DPRFS [38] is designed to cope with the large head pose variation problem in face recognition. The algorithm is depicted in Fig. 3. Before computing the face representations, a frontalized facial texture $V$ needs to be generated from the input image to normalize the pose variations [19, 38]. Let $I$ be an input image. A personalized facial model with the name AFM proposed by Kakadiaris et al. [18] is initially estimated based on the method proposed by Dou et al. [6]. The model is then fitted onto $I$ based on the landmarks detected by Wu et al. [37]. Then a frontalized facial texture $V$ (UV image) is generated through the texture lifting process proposed by Toderici et al. [33]. Along with $V$, a binary self-occlusion map $O$ is generated based on the z-buffer as in [7], where each pixel in $O$ indicates whether the corresponding pixel of $V$ in $I$ is occluded or not. In DPRFS, $V$ is divided into $n$ semi-overlapped patches as shown in Fig. 3. Let the set $P = \{P_1, P_2, P_3, ..., P_n\}$ denote the patches in
V. For each $P_i$ a separate residual network $\phi_i^R$ is learned to map $P_i$ to a face representation $x_i$, where $x_i = \phi_i^R(P_i)$. Correspondingly, $O$ is also divided into equal sized patches as $P_i$, denoted by $M_i$. A binary value $m_i$ is used to summarize the occlusion status of $M_i$. In DPRFS, if the number of visible pixel in $M_i$ is smaller than a threshold $\epsilon$, the whole patch is regarded as occluded so that $m_i = 0$. The final output of DPRFS is a concatenated face representation $x = \{x_1, x_2, ..., x_n\}$ along with the corresponding occlusion vector $m = \{m_1, m_2, ..., m_n\}$.  

3.2. Occlusion-guided compact template learning

To obtain a compact template $t$ from the face representations $x$, where the length of $t$ is at least a magnitude smaller than $x$, a mapping $\phi_C$ from the face representations to the low dimensional template space has to be created so that $t = \phi_C(x)$. Some of the previous solutions map $x$ to a low dimensional space with fully connected layers [4], PCA [32, 30, 31] or metric learning [2]. The architecture of these works is summarized in Fig. 4(L). Other solution [23] generate a compact representation $t_i$ for each individual $x_i$, with individual fully connected layers $t_i = \phi_i^C(x_i)$ and then concatenate them as $t = [t_1, t_2, ..., t_n]$. This architecture is depicted in Fig. 4(R). Assuming a facial representation $x_{i+n}$ is generated from an occluded facial patch which contains identity-irrelevant information, the corresponding template $t$ will be contaminated by the identity-irrelevant information and result in a less discriminative template. Our experiments demonstrate that the compact templates generated from these architectures are suboptimal for face recognition under large head pose variations.

The role of the occlusion mask: When a facial patch is occluded, it should have zero contribution to the final compact template. One straightforward solution to enforce this constraint is to directly convert the corresponding entries in the compact template space to zero. However, this will increase the intra-class variation when matching a frontal face to a profile face, because the distance between a zero-valued face representation generated from a profile face and a real-valued face representation from a frontal face does not reflect the actual similarity of the two faces. A possible solution proposed in DPRFS is to use $m$ in template matching. The distance $s$ between facial image $a$ and $b$ in DPRFS can be computed as:

$$s = \frac{1}{\sum(m_a \land m_b)} \sum_i \Psi_i(x^a_i, x^b_i)(m^a_i \land m^b_i),$$

where $\Psi_i(x^a_i, x^b_i)$ represents the cosine distance between two face representations $x^a_i$ and $x^b_i$. Score $s$ records the average cosine distance of all face representations computed from visible patches. The symbol $\land$ represents the logical conjunction operation.

Despite its robustness to facial occlusion, this method faces two challenges: (i) the face retrieval speed is significantly reduced by computing $n$ times the patch-wise distance (once for every pair of patches: $\Psi_i(x^a_i, x^b_i)(m^a_i \land m^b_i)$), (ii) templates generated from an image-set are very hard to be matched with $m$ because every individual sample in an image set may have a different $m$. To improve the matching speed and at the same time preserve the discriminative capability of the template in matching two image-sets, it is important to incorporate the occlusion mask in learning the compact template as $t = \phi_C(x, m)$, and the distance between image $a$ and $b$ can be computed simply by $s = \Psi(t^a, t^b)$.

Co-occurrence information in template learning: To learn a robust and discriminative feature representation for object recognition, encoding of the spatial co-occurrence among features increases the discriminative power of the final feature representation [27]. The word “contextual” is used in [8, 28] to describe the co-occurrence pattern between neighbor features and the authors claim that the contextual information encoded in facial patches enhances the robustness and stabilities of face representation. In this paper, we propose an approach that encodes the discriminative co-occurrence information in the space of $t$ using the face representations from visible patches.
It is assumed that the \( x_i \) can be mapped into a co-occurrence space using non-linear projection \( \phi^D_i \) with parameters \( \theta^D_i \) so that the face representations can be added together and create the \( t \). The reason we employ the adding operation to generate the co-occurrence information is because an adding operation will use relatively smaller space than a concatenation operation but is also able to encode the co-occurrence information from all the facial representations. To incorporate the information from the occlusion mask, \( m_i \) is extended to be a vector that has the same length as \( x \), denoted as \( m_i \). Element values in \( m_i \) simply repeat the value of \( m_i \). To compute \( t \), here are the equations:

\[
t_i = \phi^N_i (m_i, \phi^D_i(x_i; \theta^D_i); \theta^N_i),
\]

\[
t = \sum_i t_i,
\]

where non-linear mapping \( \phi^D_i \) transforms a face representation to a compact space. Mapping \( \phi^N_i \) indicates a normalization function to improve the convergence during optimization. The parameters optimized in training are \( (\theta^D_i, \theta^N_i) \). By using occlusion masks in template construction, the impact of an occluded patch to the output template would be a constant vector. The constant vector is invariant to the non-discriminative information encoded in the occluded patch since the information flow from the occluded patch is completely blocked in OGCTL. As a result, the identity irrelevant variations caused by the occluded patches will mask out the invalid information coming from \( \phi^D_i(x_i) \) and block its impact to \( t \).

Magnitude invariant loss function: Because \( t \) is a summation over \( t_i \) that come from both visible and occluded patches, the number of occluded patches in Eq. 3 may affect the magnitude of \( t \), denoted as \(|t|\). Hence, a loss function that is insensitive to \(|t|\) is required because we don’t want the similarity of two samples to depend on the number of patches that are visible. In this work, we select the recently proposed angular softmax loss (A-Softmax) [24] to learn the template embedding \( t \). Let \( j \) denote the index of an image, and \( y_j \) be the ground-truth label of the image. The loss on a training set \( \{(t_j, y_j) | j \in [1, K]\} \) that contains \( K \) samples is defined as:

\[
L = -\frac{1}{K} \sum_j \log \frac{e^{f_1(t_j, w_{y_j}, \omega)}}{\sum_c e^{f_2(t_j, w_c)}} + \sum_{c \neq y_j} e^{f_2(t_j, w_c)},
\]

\[
f_1(t_j, w_{y_j}, \omega) = ||t_j|| \cos(\omega \theta_{j,y_j}),
\]

\[
f_2(t_j, w_c) = ||t_j|| \cos(\theta_{j,c}),
\]

where symbol \( \theta_{j,y_j} \) denotes the angle between template \( t_j \) and the projection vector \( w_{y_j} \) from the \( y_j \) class. The projection vector \( w \) is learned for each class in training, and the projection vectors compose a projection matrix \( W \). The parameter \( \omega \) controls the angular margin between classes. Because the decision boundary between class \( y_j \) and class \( c(c \neq y_j) \) is only determined by \( \cos(\omega \theta_{j,y_j}) - \cos(\theta_{j,c}) \), the decision boundary of A-softmax loss is not sensitive to the magnitude of \( t \).

In summary, A-softmax loss enforces that \( t \) coming from the same class have small angular distance, while \( t \) from different classes have large angular distance, regardless the number of visible patches that contribute to \( t \). In testing, because the cosine distance is employed to compute the distance between templates, the magnitude of \( t \) will have no impact in face matching. We name the proposed algorithm Occlusion-Guided Compact Template Learning (OGCTL). The training algorithm is summarized in Alg. 1.

**Algorithm 1:** Train the neural network in OGCTL

**Input:** Training images \( \mathbb{I} \) and label vector \( y \)

**Output:** \( \{(\theta^D_i, \theta^N_i)|i \in [1, n]\} \) and \( W \)

1. **for each image \( I \) in \( \mathbb{I} \) do**
2. \[ \text{Crop facial patches } P \text{ from } I \text{ and estimate their occlusion vector } m; \]
3. \[ \text{Extract } x \text{ from } P \text{ based on } \{ \phi^D_i | i \in [1, n]\}; \]
4. \[ \text{Save } m \text{ and } x \text{ into the corresponding set } \mathbb{M} \text{ and } \mathbb{X}; \]
5. \[ \text{Optimize } \{ \theta^D_i, \theta^N_i \} \text{ in } \{ \phi^N_i, \phi^D_i \} \text{ with } \{ \mathbb{M}, \mathbb{X}, y, L \}; \]
4. Experiments

The proposed algorithms are evaluated on the UHDB31 [21] and IJB-C [25] databases. The reason that we selected these databases was because they contain large head pose variations and are still considered as challenging face recognition databases. The parameter settings of the model are illustrated in the supplementary material.

UHDB31 database [21]: This database contains 24,255 images from 77 subjects, captured under three illumination and 21 pose variations. The yaw angle variation is up to $90^\circ$ and the pitch angle variation is up to $30^\circ$. In experiments, algorithms are evaluated under all three illuminations (I03, I01, and I05) with image resolution $128 \times 153$. Abbreviation “I01” is used to denote the data partition with illumination 01. Partition “I03” is employed for evaluating the baseline models; the best-performing models are selected based on this partition to show the results and are used for the other experiments. The 1:N face identification protocol proposed by Wu et al. [55] is employed for evaluating algorithms on this database. Under this protocol, 77 frontal view images from each subject are selected to be the gallery, other images under different head pose variations are selected to be the probe images.

IJB-C database [25]: This recently released LARPA Janus Benchmarks database is a super-set of IJB-A [20] and IJB-B [35] database where all images and videos were captured from unconstrained environments. It contains 3,531 subjects with 31.3K still images and 117.5K frames from 11,779 videos. Models are evaluated on the standard 1:1 verification protocol, where a facial template is generated from a variable number of face images and video frames from different sources. In total, there are 23,124 templates with 19,557 genuine matches and 15.6M impostor matches.

4.1. Template size in face identification

The objective of the first experiment is to test methods’ performance under different sizes of output template. We change the embedding layers of deep neural networks so that the rank-1 identification performance of each method can be reported with respect to a various sizes of output templates. Multiple template generation algorithms are also compared in this experiment, which are illustrated below.

DPRFS: This baseline is the original implementation of DPRFS [38]. It employs eight ResNet-24 [12] for face recognition. As a reference, the template size of DPRFS is fixed to be the same as in the original implementations [38], which is 16 KB.

Arch.1 (AS): This baseline employs one ResNet-24 (has the same number of layers as DPRFS) to directly extract the facial template from an aligned image as [24]. A-Softmax loss (AS) is employed to train the network.

Arch.1 (S): This baseline use the same network as Arch.1 (AS) but use Softmax (S) as the loss function.

Arch.2 (AS): As depicted in Fig. 4(R), this baseline first converts face representation of DPRFS to a compact face representation with two fully connected layers (with PreLU activation function, same in Arch.3 (AS)), and then concatenates them to be the final template. It resembles the network architecture used in [23] but use AS in learning.

Arch.3 (AS): As depicted in Fig. 4(L), this baseline employs two fully-connected layers to transform the concatenated face representation generated by DPRFS into a compact template. It resembles the architecture used in [4].

In Fig. 6(T), it can be observed that OGCTL significantly outperforms architecture Arch. 1-3 in all template size variations, which demonstrates the advantages of OGCTL over other baseline models. Its performance saturates after the template size reaches 1 KB (128 dimensions floating point). Since there is only slight performance difference between the 0.5 KB and 1 KB template, in the following experiments, the 0.5 KB template generated by OGCTL is employed as the standard output of OGCTL.

Template matching speed: With a Intel 6700K CPU, the template matching speed of DPRFS is below 50K templates per second, the actual matching speed depends on the occlusion pattern of the probe image. The matching speed of OGCTL is more than 1M templates per second, it is independent from the occlusion pattern of the probe image.

4.2. Test under pose and illumination variations

The objective of the second experiment is to evaluate algorithms’ performance under pose and illumination variations. We employ all three partitions of UHDB31 database, with the notation “I03”, “I01”, and “I05”. Each partition has its own illumination condition. For each partition, the database is further divided into two sub-partitions: C15 and S6. In C15, the yaw angles of facial images are equal or below $60^\circ$. C15 contains 15 pose variations. In S6, the yaw angles of facial images are equal to $90^\circ$. So S6 contains six pose variations. The average rank-1 identification rate of every pose in each sub-partition is reported in Fig. 6(B). It is observed that the merits of DPRFS over OGCTL are mainly taken place in normal illumination condition (I03). Besides, we report the performance of Arcface [3] in Fig. 6(B), which uses a single ResNet-50 network to generate a 2KB template. It can be observed that Arcface performs better than DPRFS in C15 partition but outperforms by DPRFS in S9 partition (except I05). This experiment demonstrated that the ensemble network-based model (DPRFS) holds its merit when the pose variation is large.

To boost the performance of OGCTL further under challenging illumination and pose variations, we merge the template of DPRFS and Arcface but preserving the output template size. Specifically, every face representation $x_i$ of DPRFS is concatenated with the template of Arcface and then feed into the network of OGCTL. This method
Figure 6: Baselines’ performance with respect to: (T) different data partitions on UHDB31 database, (B) different template sizes on UHDB31 database. Note that in “S6” partitions, all the faces are partially occluded.

is named OGCTL(M). From Fig 6(B), it is observed that OGCTL(M) performs significantly better than OGCTL under large head pose variations with challenging illumination conditions. This experimental result demonstrated that the OGCTL is able to take advantage of the templates from both ensemble network model and a well-trained single network to generate a compact template for unconstrained face recognition.

4.3. Image-set based face verification

The objective of the third experiment is to further demonstrated the superior performance of OGCTL and its generalization in compact template construction. In this experiment, we evaluated the proposed algorithms under the challenging IJB-C database. The widely adopted average pooling operation [34, 38, 25] is used to merge the templates from different persons into a single template for verification. The result is shown in Table 1. Baselines were cited from the original IJB-C benchmark [25]. It can be observed from Table 1 that with a relatively small template size and a small training database, OGCTL is able to achieve improved performance than baselines that trained with a larger database with increased template size. Compared to DPRFS, because the occlusion mask is encoded in the template construction process of OGCTL, the average pooling operation only work on the discriminative co-occurrence information encoded in the visible patches, which results in a more discriminative template. Compared to OGCTL, the template of OGCTL(M) are constructed based on both DPRFS and Arcface. It outperforms Arcface in terms of AUC, achieves new state-of-the-art performance, but with only 0.5 KB template size.

4.4. Ablation study

The objective of the fourth experiment is to analyze the usefulness of each proposed module in OGCTL and disclose the unique properties of the template generated by OGCTL. The experiment is conducted on the UHDB31 database “I03” partition. We select the Pose #20 to report the face identification performance. The probe images in Pose #20 contain profile faces whose yaw variation is equal to 90°. Gallery images are all frontal faces. The template generated by OGCTL is compared to DPRFS and Arch.2 (AS). The Arch.2 (AS) is selected because it can be regarded as an learning algorithm without using the occlusion mask. We also compare an application of OGCTL without A-softmax loss but using softmax loss in learning. This baseline is denoted as OGCTL-S. The results are depicted in Table 2.

The methods are compared under two experimental settings. In the first setting, the facial templates of both gallery and probe are constructed from the three visible patches as probe. In the second setting, the templates of gallery are constructed from all the eight visible patches but the templates of probe are constructed from three visible patches. The comparison between Arch.2 (AS) and OGCTL highlights that the occlusion-guided element adding operation in OGCTL significantly improves the discriminative capabil-
Table 1: TAR (True Accept Rate) against the FAR (False positive rate) of different methods on IJB-C database. The number of images used to train each model is presented under the “scale”. The AUC represents the Area Under the ROC. The result marked with † are read from the ROC curve in the IJB-C benchmark [25]. An unavailable number is marked as “-”.

| Method              | Input          | Temp. Size | Dataset / Scale          | FAR=1E-5 | FAR=1E-4 | FAR=1E-3 | FAR=1E-2 | FAR=1E-1 | AUC     |
|---------------------|----------------|------------|--------------------------|----------|----------|----------|----------|----------|---------|
| COTS-1 [25]         | Image          | -          | -                        | 0.090    | 0.160    | 0.230    | 0.260    | 0.320    | -       |
| FaceNet [1]         | Image          | 0.5 KB     | VGGFace / 2.6M           | 0.330    | 0.490    | 0.660    | 0.820    | 0.920    | -       |
| VGG-CNN [25]        | Image          | 16 KB      | WebFace / 0.5M           | 0.430    | 0.600    | 0.750    | 0.860    | 0.950    | -       |
| Cao et al [1]       | Image          | 2 KB       | VGG2 / 3.3M              | 0.747    | 0.840    | 0.910    | 0.960    | 0.987    | -       |
| Arcface [1]         | Image          | 2 KB       | VGGFace / 2.6M           | 0.895    | 0.932    | 0.957    | 0.973    | 0.985    | 0.993   |
| DPRFS [38]          | Frontalized Img. | 16 KB  | Webface / 0.5M           | 0.310    | 0.461    | 0.638    | 0.807    | 0.939    | 0.976   |
| OGCTL               | Face representation | 0.5 KB | Webface / 0.5M           | 0.608    | 0.737    | 0.859    | 0.918    | 0.975    | 0.989   |
| OGCTL(M)            | Face representation | 0.5 KB | Webface + MS-Celeb + VGG2 / 14M | 0.879    | 0.923    | 0.952    | 0.975    | 0.988    | 0.995   |

Table 2: Performance in profile face identification. The template size marked with † are determined by the number of visible patches (depicted in Fig. 7) used for matching.

| Method | Gall. Patch Num. | Prob. Patch Num. | Temp. Size | Accuracy |
|--------|------------------|------------------|------------|----------|
| DPRFS  | 3                | 3                | 0.00 KB    | 0.961    |
| Arch2(AS) | 3                | 3                | 0.19 KB    | 0.779    |
| OGCTL-S | 3                | 3                | 0.50 KB    | 0.922    |
| OGCTL  | 3                | 3                | 0.50 KB    | 0.935    |
| OGCTL-S | 8                | 3                | 0.50 KB    | 0.883    |
| OGCTL  | 8                | 3                | 0.50 KB    | 0.974    |

Figure 7: Facial template is able to be constructed from different number of facial patches. (L) Visible patches in profile face, (R) Visible patches in frontal face.

5. Conclusion

We proposed an algorithm to construct a compact facial template based on the output of ensemble deep network models for efficient storage and fast face retrieval. The compact template integrates the face representations of visible facial patches through the proposed occlusion-guided compact template learning approach. The proposed algorithm provides a solution to generate reduced sized template with similar or better face recognition performance from ensemble networks.

References

[1] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In Proc. Automatic Face & Gesture Recognition, pages 67–74, Xi an, China, May 15-19 2018.
[2] J. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, C. Chen, V. Patel, C. Castillo, and R. Chellappa. Unconstrained still/video-based face verification with deep convolutional neural networks. International Journal of Computer Vision, 2:272 – 291, April 2018.
[3] J. Deng, J. Guo, and S. Zafeiriou. On the intrinsinc dimensionality of face representation. Arxiv preprint arXiv:1801.07698v1, 2018.
[4] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40:1002 – 1014, April 2018.
[5] Z. Dong, S. Jia, T. Wu, and M. Pei. Face video retrieval via deep learning of binary hash representations. In In Proc. AAAI Conference on Artificial Intelligence, Phoenix, AZ, February 12-17 2016.
[6] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, July 22-25 2017.
[7] P. Dou, L. Zhang, Y. Wu, S. K. Shah, and I. A. Kakadiaris. Pose-robust face signature for multi-view face recognition. In Proc. International Conference on Biometrics: Theory, Applications and Systems, pages 1–8, Arlington, VA, Sep. 8-11 2015.
[8] Y. Duan, J. Lu, J. Feng, and J. Zhou. Context-aware local binary feature learning for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5):1139–1153, May 2017.
[9] S. Gong, V. N. Boddeti, and A. K. Jain. On the intrinsic dimensionality of face representation. Arxiv preprint arXiv:1803.09672v1, 2018.
[10] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan. Deep learning with limited numerical precision. In Proc. AAAI Conference on Artificial Intelligence, 2015.
[11] W. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proc. International Conference on Computer Vision, pages 1026–1034, Santiago, Chile, December 13-16 2015.
[12] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, pages 1–15, 2016.

[13] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 142–150, Santiago, Chile, December 13-16 2015.

[14] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the Wild: A database for studying face recognition in unconstrained environments. In *Technical Report 07-49*, University of MI, Amherst, October 2007.

[15] S. Ioffe and C. Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proc. Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456, Lille, France, July 2015.

[16] C. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:117 – 128, March 2010.

[17] I. Jolliffe. *Principal component analysis*. Springer Verlag, New York, 1986.

[18] I. A. Kakadiaris, G. Passalis, G. Toderici, M. N. Mur-tuza, Y. Lu, N. Karampatziakis, and T. Theoharis. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):640–649, 2007.

[19] I. A. Kakadiaris, G. Toderici, G. Evangelopoulos, G. Passalis, D. Chu, X. Zhao, S. K. Shah, and T. Theoharis. 3D-2D face recognition with pose and illumination normalization. *Computer Vision and Image Understanding*, 154:137–151, 2017.

[20] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, Boston, MI, June 7 - 12 2015.

[21] H. Le and I. A. Kakadiaris. UHDB31: A dataset for better understanding face recognition across pose and illumination variation. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 2555–2563, Venice, Italy, October 22-29 2017.

[22] J. Lin, Z. Li, and J. Tang. Discriminative deep hashing for scalable face image retrieval. In *Proc. International Joint Conference on Artificial Intelligence*, pages 2266–2272, Melbourne, Australia, August 19 - 25 2017.

[23] J. Liu, Y. Deng, T. Bai, and C. Huang. Targeting ultimate accuracy: face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: deep hypersphere embedding for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 212 – 220, Honolulu, HI, July 21-26 2017.

[25] B. Maze, J. C. Adams, J. A. Duncan, N. D. Kalka, T. Miller, C. Otto, and A. K. Jain. Iarpa janus benchmark - c: Face dataset and protocol. In *Proc. International Conference on Biometrics*, pages 158–165, Queensland, Australia, February 20-23 2018.

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Machine Vision Conference*, pages 1–12, Swansea, UK, September 7-10 2015.

[27] X. Qi, R. Xiao, C. Li, Y. Qio, J. Guo, and X. Tang. Pairwise rotation invariant co-occurrence local binary pattern. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2199–2213, November 2014.

[28] A. Sapkota and T. Boult. Context-patch for difficult face recognition. In *Proc. International Conference on Biometrics*, pages 59–66, New Delhi, India, March 29 - April 1 2012.

[29] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. *Proc. Computer Vision and Pattern Recognition*, pages 815–823, June 2015.

[30] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Proc. Advances in Neural Information Processing Systems*, pages 1988–1996, Montreal, Canada, dec 2014.

[31] Y. Sun, D. Liang, X. Wang, and X. Tang. DeepID3: face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, pages 1–5, 2015.

[32] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1891–1898, Columbus, OH, USA, June 23-28 2014.

[33] G. Toderici, G. Passalis, S. Zafeiriou, G. Tzimiropoulos, M. Petrou, T. Theoharis, and I. A. Kakadiaris. Bidirectional relighting for 3D-aided 2D face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2721–2728, San Francisco, CA, June 13-18 2010.

[34] D. Wang, C. Otto, and A. K. Jain. Face search at scale. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1122 – 1136, June 2017.

[35] C. Whitelaam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. Iarpa janus benchmark-b face dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 592–600, Honolulu, HI, July 21-26 2017.

[36] Y. Wu, S. K. Shah, and I. A. Kakadiaris. Rendering or normalization? An analysis of the 3D-aided pose-invariant face recognition. In *Proc. IEEE International Conference on Identity, Security and Behavior Analysis*, pages 1–8, Sendai, Japan, Feb. 29-Mar. 2 2016.

[37] Y. Wu, S. K. Shah, and I. A. Kakadiaris. GoDP: globally optimized dual pathway system for facial landmark localization in-the-wild. *Image and Vision Computing*, 73:1–16, 2018.

[38] X. Xu, H. Le, P. Dou, Y. Wu, and I. A. Kakadiaris. Evaluation of 3D-aided pose invariant 2D face recognition system. In *Proc. International Joint Conference on Biometrics*, pages 446–455, Denver, Colorado, October 2017.