A Good Practice Towards Top Performance of Face Recognition: Transferred Deep Feature Fusion

Lin Xiong[1*†], Jayashree Karlekar[1*], Jian Zhao[2*†], Jiashi Feng[2], *Member, IEEE*, and Shengmei Shen[1]

***Abstract*—Unconstrained face recognition performance evaluations have traditionally focused on Labeled Faces in the Wild (LFW) dataset for imagery and the YouTubeFaces (YTF) dataset for videos in the last couple of years. Spectacular progress in this field has resulted in a saturation on verification and identification accuracies for those benchmark datasets. In this paper, we propose a unified learning framework named transferred deep feature fusion targeting at the new IARPA Janus Bechmark A (IJB-A) face recognition dataset released by NIST face challenge. The IJB-A dataset includes real-world unconstrained faces from 500 subjects with full pose and illumination variations which are much harder than the LFW and YTF datasets. Inspired by transfer learning, we train two advanced deep convolutional neural networks (DCNN) with two different large datasets in source domain, respectively. By exploring the complementarity of two distinct DCNNs, deep feature fusion is utilized after feature extraction in target domain. Then, template specific linear SVMs is adopted to enhance the discrimination of framework. Finally, multiple matching scores corresponding different templates are merged as the final results. This simple unified framework outperforms the state-of-the-art by a wide margin on IJB-A dataset. Based on the proposed approach, we have submitted our IJB-A results to National Institute of Standards and Technology (NIST) for official evaluation.**

***Index Terms*—Face Recognition, Deep Convolutional Neural Network, Feature Fusion, Model Ensemble, SVMs.**

I. Introduction

FACE recognition performance using features of Deep Convolutional Neural Network (DCNN) have been dramatically improved in recent years. Many state-of-the-art algorithms claim very close [1],[2] or even have surpassed [3], [4],[5] human performance on Labeled Faces in the Wild (LFW) dataset. The saturation in recognition accuracy for current benchmark dataset has come. In order to push the development of frontier in regarding to unconstrained face recognition, a new face dataset template-based IJB-A is introduced recently [6], whose setting and solutions are aligned better with the requirements of real applications.

The IJB-A dataset is created to provide the latest and most challenging dataset for both verification and identification as shown is Fig.1. Unlike LFW and YTF, this dataset includes both image and video of subjects manually annotated with facial bounding boxes to avoid the near frontal condition, along with protocols for evaluation of both verification and identification. Those protocols significantly deviate from standard protocols for many face recognition algorithms [7],[8]. Moreover, the concept of template is introduced, simultaneously. A template refers to a collection of all media (images and/or video frames) of an interested face captured under different conditions that can be utilized as a combined single representation for matching task. The template-based setting reflects many real-world biometric scenarios, where capturing a subject's facial appearance is possible more than once under different acquisition ways. In other words, this new IJB-A face recognition task requires to deal with a more challenging set-to-set matching problem successfully regardless of face capture settings (illumination, sensor, resolution) or subject conditions (facial pose, expression, occlusion).

(a) Face recognition over single image.

(b) Unconstrained set-based face recognition.

Fig. 1: Comparison between face recognition over single image and unconstrained set-based face recognition. (a) Face recognition over single image. (b) Unconstrained set-based face recognition where each subject is represented by a set of mixed images and videos captured under unconstrained conditions. Each set contains large variations in face pose, expression, illumination and occlusion issues. Existing single-medium based recognition approaches cannot successfully address this problem consistently. Matched cases are bounded with green boxes, while non-matched cases are bounded with red boxes. Best viewed in color.

Our contributions can be summarized as following aspects:

1) A unified learning framework named transferred deep feature fusion is proposed for face verification and identification.
2) Two latest DCNN models are trained in source domain with two different large datasets in order to take full advantage of complementary between models and datasets.
3) Two-stage fusion are designed, one for features and

[1]L. Xiong, J. Karlekar and S.M. Shen are with Core Technology Group, Learning & Vision, Panasonic R&D Center Singapore, Singapore (lin.xiong@sg.panasonic.com; karlekar.jayashree@sg.panasonic.com; shengmei.shen@sg.panasonic.com).

[2]J. Zhao and J.S. Feng are with Department of Electrical and Computer Engineering, National University of Singapore, Singapore (zhaojian90@u.nus.edu; elefjia@nus.edu.sg). J. Zhao was an intern at Core Technology Group, Learning & Vision, Panasonic R&D Center Singapore during this work.

* L. Xiong, J. Zhao and J. Karlekar make an equal contribution.

† L. Xiong and J. Zhao are the corresponding author.

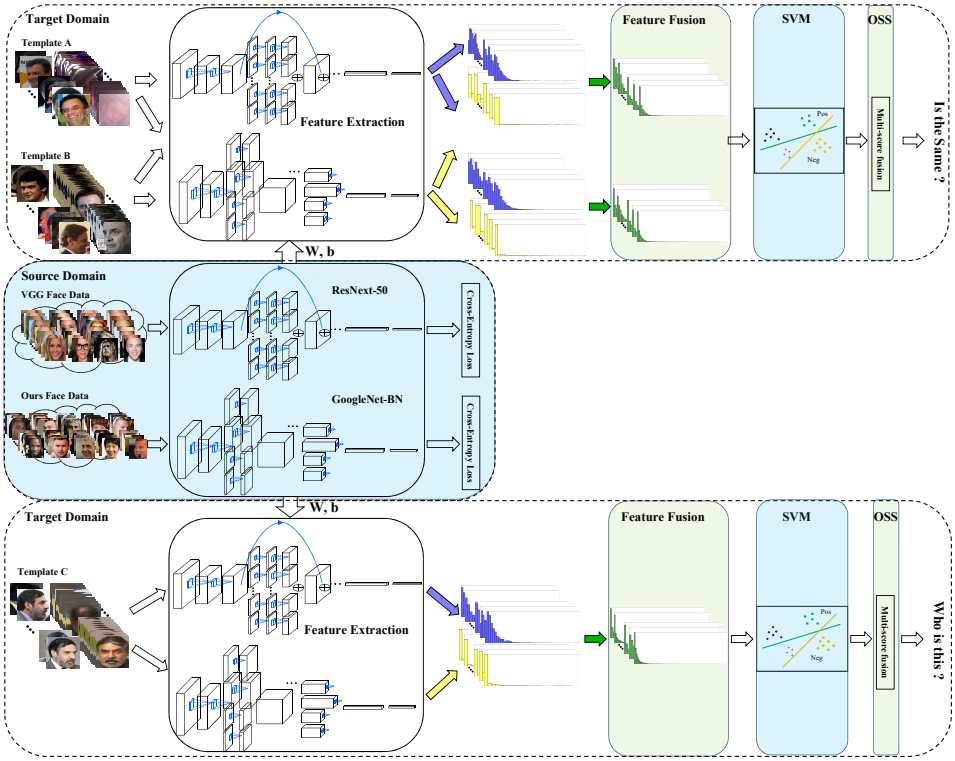

Fig. 2: Framework overview. Our learning framework consists three components: Deep feature learning module locates middle component, Template-based unconstrained face recognition is included in upper and lower components. Training procedures are illustrated with blue blocks, two-stage fusion is depicted in green blocks. Best viewed in color.

another for similarity scores.

4) One-vs-rest template specific linear SVMs with chosen negative set is trained in target domain.

In this paper, we propose a unified learning framework named transferred deep feature fusion. It can effectively integrate superiority of each module and outperform the state-of-the-art on IJB-A dataset. Inspired by transfer learning [9], facial feature encoding model of subjects are trained offline in a source domain, and this feature encoding model is transferred to a specific target domain where limited available faces of new subjects can be encoded. Specifically, in order to capture the intrinsic discrimination of subjects and enhance the generalization capability of face recognition models, we deploy two advanced deep convolutional neural networks (DCNN) with distinct architectures to learn the representation of faces on two different large datasets (each one has no overlap with IJB-A dataset) in source domain. These two DCNN models provide distinct feature representations which can better characterize the data distribution from different perspectives. The complementary between two distinct models is beneficial for feature representation [10]. Thus, representing a face from different perspectives could effectively decrease ambiguity among subjects and enhance the generalization performance of face recognition especially on extremely large number of subjects. After offline training procedure, those two DCNN models are transferred to target domain where templates of IJB-A dataset as inputs are performed feature extraction with shared weigths and biases, respectively. Then, features from two DCNN models are combined in order to obtain more discriminative representation. Finally, template specific linear SVMs are trained on fused features for classification. Furthermore, for set-to-set matching problem, multiple matching scores are merged into a single one [11],[12],[13] for each template pair as the final results. Comprehensive evaluations on IJB-A public dataset well demonstrate the significant superiority of the proposed learning framework over other state-of-the-art methods. Based on the proposed approach, we have submitted our IJB-A results to NIST for official evaluation.

This paper is organized as follows. We review the related work in Section II. Section III shows the details of transferred deep feature fusion. In Section IV, a comprehensive evaluation on IJB-A dataset is shown. Finally, the conclusion remarks and the future work are presented in Section V.

II. Related Work

Recently, all the top performing methods for face recognition on LFW and YTF are all based on DCNN architectures. Such as the VGG-Face model [14], as a typical application of the VGG-16 convolutional network architecture [15] trained on a reasonably and publicly large face dataset of 2.6M images of 2622 subjects, provides state-of-the-art performance. This dataset is called as VGG-Face data for convenience in the following section. FaceNet [4] utilizes the DCNN with inception module [16] for unconstrained face recognition. This network is trained using a private huge dataset of over 200M images and 8M subjects. DeepFace [1] deploys a DCNN coupled with 3D alignment, where facial pose is normalized by warping facial landmarks to a canonical position prior to encoding face images. DeepID2+ [2] and DeepID3 [3] extend the FaceNet model by including joint Bayesian metric learning [17] and multi-task learning. More better unconstrained face recognition performance is provided by them. Moreover, DeepFace is trained using a private dataset of 4.4M images and 4,030 subjects. DeepID2+ and DeepID3 are trained also using a private dataset of 202,595 images and 10,117 subjects with 25 networks and 50 networks, respectively. The idea of multiple model ensemble is involved. Moreover, many approaches use metric learning in the form of triplet loss similarity or joint Bayesian for the final loss to learn an optimal embedding for face recognition [4],[14],[5]. Thus, a recent study [18] concludes that multiple networks ensemble and metric learning are crucial for improvement on LFW. With the advent of IJB-A dataset introduced by

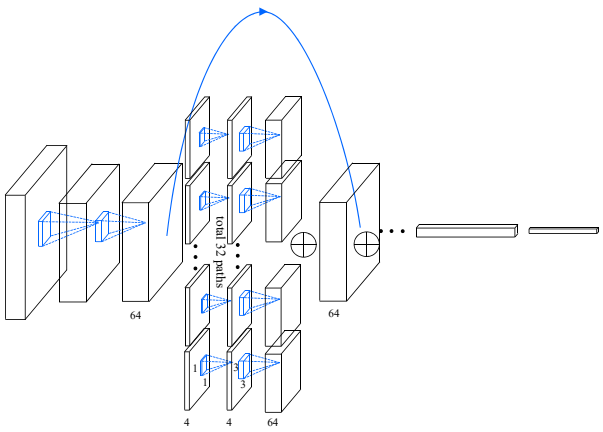

Fig. 3: A block of ResNext with cardinality=32.

NIST in 2015, the task of template-based unconstrained face recognition has attracted extensive attention. So far as we known, most algorithms for this challenging problem are also based on DCNN architecture as top performing methods did on LFW and YTF. Chen *et al.* [5] achieve good performance by extracting feature representations via a DCNN trained on public dataset which includes 490,356 images and 10,548 subjects. And then, those features as inputs are applied to learn metric matrix in order to project the feature vector into a low-dimensional space, meanwhile, maximizing the between-class variation and minimizing within-class variation via joint Bayesian metric learning. B-CNN [19] applies the bilinear CNN architecture to face identification. Deep Multi-pose [20] utilizes five pose specialized sub-networks with 3D pose rendering to encode multiple pose-specific features. Sensitivity of the recognition system to pose variations is reduced since an ensemble of pose-specific deep features is adopted. Pooling faces [12] aligns faces in 3D and bins them according to head pose and image quality. Pose-Aware Models (PAMs) [11] handles pose variability by learning Pose-Aware Models for frontal, half-profile and full-profile poses in order to improve face recognition performance in wild. Masi *et al.* [13] even question whether need to collect millions of faces or not for effective face recognition. Thus, a far more accessible means of increasing training data sizes is proposed. Pose, 3D shape and expression are utilized to synthesize more faces from CASIA-WebFace dataset [21]. Triplet Probabilistic Embedding (TPE) [22] couples a DCNN-based approach with a low-dimensional discriminative embedding learned using triplet probability constraints to solve the unconstrained face verification problem. TPE obtains better performance than previous algorithms on IJB-A dataset. Template Adaptation (TA) [23] proposes the idea of template adaptation which is a form of transfer learning to the set of media in a template. Combining DCNN features with template adaptation, it obtains better performance than TPE on IJB-A task. Ranjan *et al.* propose an all-in-one method [24] employed a multi-task learning framework that regularizes the shared parameters of CNN and builds a synergy among different domains and tasks. Until recently, Yang *et al.* propose Neural Aggregation Network (NAN) [25] which produces a compact and fixed-dimension feature representation. It adaptively aggregates the features to form a single feature inside the convex hull spanned by them. What's more interesting is that NAN learns to advocate high-quality face images while repelling low-quality ones such as blurred, occluded and improperly exposed faces. Thus, the face recognition performance on IJB-A dataset is pushed to reach an unprecedented height. Just a few days ago, Ranjan *et al.* add an L_2-constraint to the feature descriptors which restricts them to lie on a hypersphere of a fixed radius. Therefore, minimizing the softmax loss is equivalent to maximizing the cosine similarity for the positive pairs and minimizing it for the negative pairs. In this way, the verification performance on IJB-A dataset is refreshed again.

In the current work, we also follow the similar way– DCNN model should be a good baseline. By virtue of the complementary between different DCNN architectures and datasets, we can obtain a more general feature representation model via ensemble strategy. Intrinsic discrimination of subjects is also important for face recognition, inspired by transfer learning, template specific linear one-vs-rest SVMs are trained in target domain. It shares similar idea as TA [23] while different negative set is chosen. Similar to [11],[12],[13], multiple matching scores are merged into a single one for set-to-set matching whereas an easier way is adopted. Last, we also deploy TPE to further enhance performance of face

recognition. More detailed information about our learning framework can be found in the next section part.

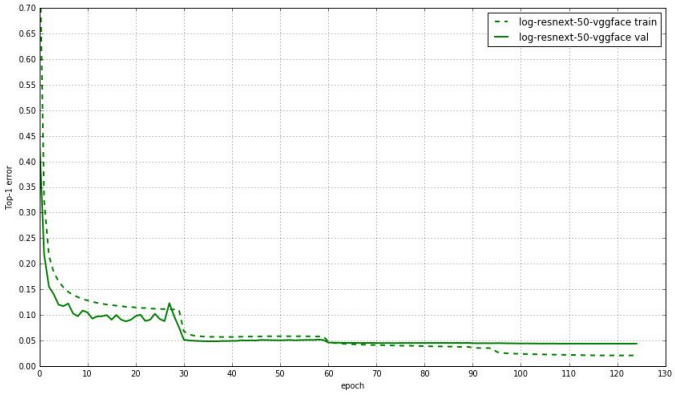

Fig. 4: Training on VGG-Face data. Solid curve denotes top 1 training error, and dotted line denotes validation error of the center crops.

III. Transferred Deep Feature Fusion

It is necessary that DCNN architectures are trained on tremendous dataset. However, IJB-A datasets contains 500 subjects with 5,396 images and 2,042 videos sampled to 20,412 frames in total. This is obviously inadequate. Unlike [13] where training data is increased by synthesizing faces based on pose, 3D shape and expression variations, inspired by domain adaptation, we need other huge labeled face datasets in source domain to train DCNN model. It is different from replacing the final entropy loss layer for a new task and fine-tuning the DCNN model on this new objective using data from the target domain [26]. We focus on training DCNN model and the one-vs-rest linear SVMs in source domain and target domain, separately. Last, one-shot-similarity (OSS) [27] is utilized to calculate similarity scores and we fuse those multiple matching scores into a single one for final performance evaluation. As shown in Fig.2, our learning framework consists three components: two distinct DCNN models are trained with two different large face datasets in source domain illustrated in middle component, respectively. In target domain, the new unseen data as inputs are fed into those two DCNN architectures with the shared weights and biases learned from source domain for feature extraction, respectively. Then, all features are combined in the first fusion stage. Template specific one-vs-rest SVMs are trained on those fused features in order to boost the intrinsic discrimination of subjects. Last but not least, multiple matching scores computed by OSS is weighted to one final score for verification and identification in the second fusion stage of upper and lower components, respectively. The detailed of each components of our learning framework are presented in the following subsections.

A. Deep feature learning in source domain

In this part, we discuss detailedly two DCNN models and two extra huge datasets for training in source domain.

Since Network-in-Network (NIN) [28] has been proposed, the depth of DCNN is refreshed again and again. Recent works [29],[30],[31] have shown that convolutional networks with small filters can be substantially deeper, more accurate, and efficient to train if they contain shorter connections between layers close to the input and those close to the output. The bypassing paths are presumed to be the key factor that eases the training of these very deep networks. This point is further supported by ResNets [32], in which pure identity mappings are used as bypassing paths. ResNets have achieved impressive, record-breaking performance on ImageNet [33]. Until recently, Xie *et al.* [34] reconstruct the building block of ResNets with aggregating a set of transformations. This simple design results in a homogeneous, multi-branch architecture that has only a few hyper-parameters to set. A new dimension called *cardinality* is proposed, which as an essential factor in addition to the dimension of depth and width. Thus, it is codenamed ResNext. A typical block of ResNext is shown in Fig.3. Considering the balance between performance and efficiency, we choose ResNext 50 as the first DCNN model.

For public large face dataset, the VGG-Face should be a better choice for ResNext 50. The original VGG-Face dataset includes 2,109,307 available images and 2,614 subjects. First, we utilize ground-truth bounding box given by dataset to crop and resize face images from the original ones. Each face image is 144×144. An off-the-shelf CNN model pre-trained on CASIA-WebFace is deployed to do noisy data cleaning. Moreover, the overlap subject with IJB-A dataset should be removed. Finally, we obtain 1,648,187 images and 2,613 subjects in total. For partition of training and validation parts, we refer to ImageNet. 90% of the total images (1,483,368) are served as training data. 5% of the total images (82,410) are viewed as validation data. Our implementation for VGG-Face on ResNext 50 is implemented by MXNet [35]. The image is resized from 144×144 to 480×480 for data augmentation. A 224×224 crop is randomly sampled from 480×480 or its horizontal flip, with the per-pixel mean substracted. The standard color augmentation [36] is used. We adopt batch normalization (BN) [37] right after each convolution and before ReLU. We initialize the weights as in [38] and train ResNext 50 from scratch. NAG with a mini-batch size of 256 is utilized on 4 GPUs (NVIDIA M40 with 12 GB GDDR) . The learning rate starts from 0.1 and is divided by 10 every 30 epoch and the model is trained for up to 125 epoch. The weight decay is 0.0001 and the momentum is 0.9. The cardinality is 32. The training and validation curves are shown in Fig.4. Finally, we obtain the validation performance 95.63% at top1 and 97.00% at top 5, respectively.

Inspired by NIN, an orthogonal approach to making networks deeper (e.g., with the help of skip connections) is to increase the network width. The GoogLeNet [16] uses an "Inception module" which concatenates features maps produced by filers of different sizes. Different from ResNext which enhances representational power of network via extremely deep architecture, GoogLeNet depends on wider structure to boost capacity of network. Along with the BN emergence, training DCNN becomes easier than before. Thus, GoogLeNet-BN is our second DCNN model.

To train GoogLeNet-BN, we collect around 8M faces images of 99K subjects via crawling from the Internet. Data

Fig. 5: Face identification results for IJB-A split1 on close protocol. The first column shows the query images from probe templates. The remaining 5 columns show the corresponding top-5 queried gallery templates.

preprocessing is done as following steps. We use OpenCV to detect face and utilize bounding box to crop and resize face images. Each image is 256×256. There are 582,405 images can not be detected, so we delete them. The overlap subject with IJB-A dataset should be removed. Considering the data distribution, we only keep those identities which have 40-500 images. Finally, we obtain 4,356,052 images and 53,317 subjects in total. Our implementation for our face data on GoogLeNet-BN is implemented by caffe [39]. A 224×224 crop is randomly sampled from 256×256 or its horizontal flip. We initialize the weights as in [38] and train GoogLeNet from scratch. SGD with a mini-batch size of 256 is utilized on 4 GPUs (NVIDIA M40 with 12 GB GDDR). The learning rate starts from 0.1 and exp policy is adopted. The weight decay is 0.0001 and the momentum is 0.9. The model are trained for up to 60×10^4 iterations. We stop training procedure when the error is not decreasing.

B. Template-based unconstrained face recognition

After finish training procedure of two DCNN models in source domain. Weights and biases of ResNext 50 and GoogLeNet-BN are shared into target domain. Each face image or frame of video is viewed as input to feed into those two models, respectively. For ResNext 50, the penultimate global average pooling layer is served as feature extraction layer. It has 2,048 output size. Thus, the feature dimension is 2,048. Given an image or frame $\mathbf{x}_i \in \mathbb{R}^d$ from a mini-batch of size M, $f_R(\mathbf{x}_i) \in \mathbb{R}^{d_1}$ denotes the feature from ResNext 50, where $d_1 < d$ and $d_1 = 2048$. Similarly, for GoogLeNet-BN, 7×7 average pooling layer is treated as feature extraction layer. The channel size is 1,024. So, the feature dimension is 1,024. Let $f_G(\mathbf{x}_i) \in \mathbb{R}^{d_2}$ is the feature from GoogLenet-BN, where $d_2 = 1024$. In the first-stage fusion, $f_R(\mathbf{x}_i)$ and $f_G(\mathbf{x}_i)$ are concatenated into $f_F(\mathbf{x}_i) \in \mathbb{R}^{d_3}$, where $d_3 = 3072$.

After feature fusion, in order to train a more discriminative

TABLE I: Performance evaluation on the IJB-A dataset. For 1:1 verification, the true accept rates (TAR) @ false positive rates (FAR) are presented. For 1:N identification, the true positive identification rate (TPIR) @ false positive identification rate (FPIR) and CMC are reported

Method	1:1 Verification TAR			1:N Identification TPIR				
	FAR=0.001	FAR=0.01	FAR=0.1	FPIR=0.01	FPIR=0.1	Rank 1	Rank 5	Rank 10
OpenBR[40]	0.104±0.014	0.236±0.009	0.433±0.006	0.066±0.017	0.149±0.028	0.246±0.011	0.375±0.008	-
GOTS[6]	0.198±0.008	0.406±0.014	0.627±0.012	0.047±0.024	0.235±0.033	0.433±0.021	0.595±0.020	-
B-CNN[19]	-	-	-	0.143±0.027	0.341±0.032	0.588±0.020	0.796±0.017	-
Pooling faces[12]	-	0.309	0.631	-	-	0.846	0.933	0.951
LSFS[41]	0.514±0.060	0.733±0.034	0.895±0.013	0.383±0.063	0.613±0.032	0.820±0.024	0.929±0.013	-
Deep Multi-pose[20]	-	0.787	0.911	0.52	0.75	0.846	0.927	0.947
DCNN*manual*+metric[42]	-	0.787±0.043	0.947±0.011	-	-	0.852±0.018	0.937±0.010	0.954±0.007
Triplet Similarity[43]	0.590±0.050	0.790±0.030	0.945±0.002	0.556±0.065	0.754±0.014	0.880±0.015	0.950±0.007	0.974±0.006
VGG-Face[14]	-	0.805±0.030	-	0.461±0.077	0.670±0.031	0.913±0.011	-	0.981±0.005
PAMs[11]	0.652±0.037	0.826±0.018	-	-	-	0.840±0.012	0.925±0.008	0.946±0.007
DCNN*fusion*[5]	-	0.838±0.042	0.967±0.009	0.577±0.094	0.790±0.033	0.903±0.012	0.965±0.008	0.977±0.007
Masi *et al.*[13]	0.725	0.886	-	-	-	0.906	0.962	0.977
Triplet Embedding[22]	0.813±0.020	0.900±0.010	0.964±0.005	0.753±0.030	0.863±0.014	0.932±0.010	-	0.977±0.005
Template Adaptation[23]	0.836±0.027	0.939±0.013	0.979±0.004	0.774±0.049	0.882±0.016	0.928±0.010	0.977±0.004	0.986±0.003
All-In-One+TPE[24]	0.823±0.020	0.922±0.010	0.976±0.004	0.792±0.020	0.887±0.014	0.947±0.008	-	0.988±0.003
NAN[25]	0.881±0.011	0.941±0.008	0.978±0.003	0.817±0.041	0.917±0.009	0.958±0.005	0.980±0.005	0.986±0.003
L_2-softmax[44]	0.906±0.016	0.952±0.007	0.981±0.003	0.852±0.042	0.930±0.010	0.963±0.007	-	0.986±0.002
L_2-softmax[44]+TPE[22]	0.910±0.013	0.951±0.006	0.979±0.003	0.873±0.024	0.931±0.010	0.961±0.007	-	0.983±0.003
TDFF	0.919±0.006	**0.961±0.007**	0.988±0.003	0.878±0.035	**0.941±0.010**	**0.964±0.006**	**0.988±0.003**	**0.992±0.002**
TDFF+TPE[22]	**0.921±0.005**	**0.961±0.007**	**0.989±0.003**	**0.881±0.039**	**0.940±0.009**	**0.964±0.007**	**0.988±0.003**	**0.992±0.003**

model in target domain, template specific one-vs-rest SVMs play an important role. Specifically, the weights and biases terms for template specific SVMs are learned by optimizing the following L_2-regularized L_2-loss objective function:

$$\begin{aligned}\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \lambda_+ \sum_{i=1}^{N_+} \max\left[0, 1 - y_i \mathbf{w}^T f_F\left(\mathbf{x}_i\right)\right]^2 \\ +\lambda_- \sum_{i=1}^{N_-} \max\left[0, 1 - y_j \mathbf{w}^T f_F\left(\mathbf{x}_j\right)\right]^2\end{aligned} \tag{1}$$

where $\mathbf{w}$ denote the weights including bias term, $y_i \in \{-1, 1\}$ denotes the label indicating whether the current sample being negative or possible, N_+ indicates the number of positive samples, N_- is the number of negative ones, $N_- \gg N_+$. Moreover, the constraint for negative samples $\lambda_- = C\frac{N_+ + N_-}{2N_-}$, the constraint for positive samples $\lambda_+ = C\frac{N_+ + N_-}{2N_+}$, where C is a trade-off factor. A template includes images or/and frames of video. For the feature of video frame, we compute the average media encodings. Let t_j^V denotes average media encoding of video j.

$$t_j^V = \frac{1}{N_j^V}\sum_{i=1}^{N_j^V} f_F\left(\mathbf{x}_i\right) \tag{2}$$

where N_j^V is the number of frame in video j, $\mathbf{x}_i$ denotes i frame of video j. In other words, all features of video frames are aggregate one feature. Thus, the deep facial representations for the ath template can be expressed as

$$T_a = \left\{t_i^I, ..., t_{N_a}^V\right\} \tag{3}$$

where t_i^I denotes ith image, N_a express the number of image and video. All media encoding need to perform unit normalization. For verification (a.k.a 1:1 compare), the positive sample of template specific SVM is probe template, the large-scale negative samples include the whole training set. For identification (a.k.a 1:N search), the probe template specific SVMs adopt the whole training set as the large-scale negative samples; whereas for gallery template specific SVM, we adopt other gallery templates and the whole training set as large-scale negative samples. Based on One shot similarity (OSS), we compute similarity between two features p and q via $s\left(p, q\right) = \frac{1}{2}\mathcal{P}\left(q\right) + \frac{1}{2}\mathcal{Q}\left(p\right)$ where $\mathcal{P}\left(q\right)$ denotes the trained probe template specific SVM model and $\mathcal{Q}\left(p\right)$ indicates the trained gallery template specific SVM model. One template exists many features as Eqn.3, the resulting multiple matching scores should be ensembled into a single one for each template pair in second-stage fusion.

$$s\left(T_a, T_b\right) = \frac{\sum\limits_{t_i \in T_a, t_j \in T_b} s\left(t_i, t_j\right) e^{\beta\ s(t_i, t_j)}}{\sum\limits_{t_i \in T_a, t_j \in T_b} e^{\beta\ s(t_i, t_j)}} \tag{4}$$

where $\beta = 0$ is enough in our following experiments.

TABLE II: Performance evaluation on the IJB-A dataset. For 1:1 verification, the true accept rates (TAR) @ false positive rates (FAR) are presented.

Method	1:1 Verification TAR
	FAR=0.0001
L_2-softmax[44]	0.832±0.027
L_2-softmax[44]+TPE[22]	0.863±0.012
TDFF	**0.875±0.013**
TDFF+TPE[22]	**0.877±0.018**

IV. Experiments and analysis

In this section, we describe the results for evaluation of the experimental system on the IJB-A verification and identification protocols. The IJB-A dataset contains face images and video frames captured from unconstrained settings which are aligned better with the requirements of real applications. There are 500 subjects with 5,396 images and 2,042 videos

(a) The best mated template pairs (b) The worst mated template pairs

Fig. 6: Verification results analysis for mated template pairs on IJB-A split1 .

sampled to 20,412 frames in total. Full pose variation and wide variations in imaging conditions are the main features of IJB-A dataset, which makes the face recognition very challenging. In our experiments, we just utilize the ground-truth bounding box to crop face image from the original one and resize to 224×224 for each image or frame. We do not use any off-the-shelf pre-trained DCNN model to clean data. We also do not deploy any face detector and do not perform any face alignment procedure.

A remarkable feature of this dataset is that the concept of template is introduced. Each training an testing sample in called a template which comprises a mixture of static images and sampled video frames. Each static image or a frame of video corresponds with a media. On average, each subject has 11.4 images and 4.2 videos. There are 10 training and testing splits. Each of them contains 333 and 167 subjects, respectively.

In TableI, we list the performance of state-of-the-art algorithms on IJB-A dataset. Our performance achieves the best of them for both verification and identification protocols. When we use the TPE to learn a discriminative mapping space while keep the original feature dimension using the training splits of IJB-A. It slightly improves the performance and achieves the new record TAR of 0.921 @ FAR = 0.001, TAR of 0.961 @ FAR = 0.01 and TAR of 0.989 @ FAR = 0.1 for verification. Our method performs significantly better than state-of-the-other algorithms in other indicators as well. These results clearly suggests the effectiveness of our proposed learning framework. In [44], the author reports the results for a very low FAR of 0.0001. Thus, in TableII, we also report the performance @ FAR = 0.0001 for verification protocol, our results still slightly better than L_2-softmax, even TPE is added.

We illustrate the identification results for IJB-A split1 on close protocol in Fig.5. The first column shows the query images from probe templates. The remaining 5 columns show the corresponding top-5 queried gallery templates. For each template, we provide Template ID, Subject ID and similarity score. For all five rows, our approach can successfully find the subjects in rank 1.

Finally, we visualize the verification results in Fig.6 and Fig.7 for IJB-A split1 to gain insight into template based unconstrained face recognition. After computing the similarities for all pairs of probe and reference templates, we sort the resulting list. Each row represents a probe and reference template pair. The original templates within IJB-A contain from one to dozens of media. Up to eight individual media are shown with the last space showing a mosaic of the remaining media in the template. Between the templates are the Template

(a) The best nonmated template pairs (b) The worst nonmated template pairs

Fig. 7: Verification results analysis for nonmated template pairs on IJB-A split1 .

IDs for probe and reference as well as the best mated and best non-mated similarity. Fig.6 (a) shows the highest mated similarities. In the thirty highest scoring correct matches, we note that every reference template contains dozens of media. The probe templates also contain dozens of media that matches well. Fig.6 (b) shows the lowest mated template pairs, representing failed matching. The thirty lowest mated results from single-media reference templates are under extremely challenging unconstrained conditions. There extremely difficult cases cannot be solved even using our proposed approach. Fig.7 (a) showing the best non-mated similarities shows the most certain nonmates, again often involving large templates with enough guidance from the relevant and historical information. Fig.7 (b) showing the worst non-mated pairs highlights the unstable errors involving single-media reference templates representing impostors in challenging orientation.

V. Conclusion

In this paper, we propose a unified learning framework named transferred deep feature fusion. It can effectively integrate superiority of each module and outperform the state-of-the-art on IJB-A dataset. Inspired by transfer learning, facial feature encoding model of subjects are trained offline in a source domain, and this feature encoding model is transferred to a specific target domain where limited available faces of new subjects can be encoded. Specifically, in order to capture the intrinsic discrimination of subjects and enhance the generalization capability of face recognition models, we deploy two advanced deep convolutional neural networks (DCNN) with distinct architectures to learn the representation of faces on two different large datasets (each one has no overlap with IJB-A dataset) in source domain. These two DCNN models provide distinct feature representations which can better characterize the data distribution from different perspectives. The complementary between two distinct models is beneficial for feature representation. Thus, representing a face from different perspectives could effectively decrease ambiguity among subjects and enhance the generalization performance of face recognition especially on extremely large number of subjects. After offline training procedure, those two DCNN models are transferred to target domain where templates of IJB-A dataset as inputs are performed feature extraction with shared weigths and biases, respectively. Then, two-stage fusion is designed, features from two DCNN models are combined in order to obtain more discriminative representation in first-stage. Finally, template specific linear SVMs are trained on fused features for classification. Furthermore, for set-to-set matching problem, multiple matching scores are merged into

a single one for each template pair as the final results in the second-stage of fusion. Comprehensive evaluations on IJB-A public dataset well demonstrate the significant superiority of the proposed learning framework over other state-of-the-art methods. Based on the proposed approach, we have submitted our IJB-A results to NIST for official evaluation. In the feature, end-to-end network architecture is still attractive for face recognition. Manifold-based metric learning can learn non-linear embedding space, it can explore the geometric structure of the feature encoding.Because, the rotation of head follows a low-dimension manifold. Dictionary learning combines DCNN is an interesting task.

REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[2] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900.

[3] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[5] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[6] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1931–1939.

[7] H. Ye, W. Shao, H. Wang, J. Ma, L. Wang, Y. Zheng, and X. Xue, "Face recognition via active annotation and learning," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1058–1062.

[8] J. Li, J. Zhao, F. Zhao, H. Liu, J. Li, S. Shen, J. Feng, and T. Sim, "Robust face recognition with deep multi-view representation learning," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1068–1072.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[10] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.

[11] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4838–4846.

[12] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni, "Pooling faces: template based face recognition with pooled face images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 59–67.

[13] I. Masi, A. T. Trá°§n, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *European Conference on Computer Vision*. Springer, 2016, pp. 579–596.

[14] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[17] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*. Springer, 2012, pp. 566–579.

[18] 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 *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 142–150.

[19] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear cnns," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[20] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan *et al.*, "Face recognition using deep multi-pose representations," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[21] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[22] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*. IEEE, 2016, pp. 1–8.

[23] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," *arXiv preprint arXiv:1603.03958*, 2016.

[24] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," *arXiv preprint arXiv:1611.00851*, 2016.

[25] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," *arXiv preprint arXiv:1603.05474*, 2016.

[26] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[27] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 10, pp. 1978–1990, 2011.

[28] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[29] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377–2385.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[31] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[34] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *arXiv preprint arXiv:1611.05431*, 2016.

[35] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[39] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[40] J. C. Klontz, B. F. Klare, S. Klum, A. K. Jain, and M. J. Burge, "Open source biometric recognition," in *Biometrics: Theory, Applications and*

Systems (BTAS), 2013 IEEE Sixth International Conference on. IEEE, 2013, pp. 1–8.
[41] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," *arXiv preprint arXiv:1507.07242*, 2015.
[42] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa, "An end-to-end system for unconstrained face verification with deep convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 118–126.
[43] S. Sankaranarayanan, A. Alavi, and R. Chellappa, "Triplet similarity embedding for face verification," *arXiv preprint arXiv:1602.03418*, 2016.
[44] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507v1*, 2017.

Lin Xiong received the B.S. degree from Shaanxi University of Science & Technology in 2003, and he received the Ph.D. degree with School of Electronic Engineering, Xidian University, China, in 2014. He is currently an research engineer of Learning & Vision, Core Technology Group, Panasonic R&D Center Singapore, Singapore. His current research interests include face recognition, person re-identification, deep learning, Riemannian manifold optimization, low-rank and sparse matrix factorization, background modeling.

PLACE PHOTO HERE

Jayashree Karlekar

Jian Zhao received the B.S. degree from Beihang University in 2012, and he received the Master degree with School of Computer, National University of Defense Technology, China, in 2014. He is currently funded by China Scholarship Council (CSC) and School of Computer, National University of Defense Technology to pursue his Ph.D. degree at Learning and Vision Group, Department of Electronical and Computer Engineering, Faculty of Engineering, National University of Singapore. His current research interests include face recognition, human parsing, human pose estimation, object detection, object semantic segmentation, and relevant deep learning and computer vision problems.

Jiashi Feng is currently an assistant Professor in the department of electrical and computer engineering in the National University of Singapore. He got his B.E. degree from University of Science and Technology, China in 2007 and Ph.D. degree from National University of Singapore in 2014. He was a postdoc researcher at University of California from 2014 to 2015. His current research interest focus on machine learning and computer vision techniques for large-scale data analysis. Specifically, he has done work in object recognition, deep learning, machine learning, highdimensional statistics and big data analysis.

PLACE PHOTO HERE

Shengmei Shen