Robust Face Recognition via Multimodal Deep Face Representation

Changxing Ding, *Student Member, IEEE*, Dacheng Tao, *Fellow, IEEE*

arXiv:1509.00244v1 [cs.CV] 1 Sep 2015

Abstract—Face images appeared in multimedia applications, *e.g.*, social networks and digital entertainment, usually exhibit dramatic pose, illumination, and expression variations, resulting in considerable performance degradation for traditional face recognition algorithms. This paper proposes a comprehensive deep learning framework to jointly learn face representation using multimodal information. The proposed deep learning structure is composed of a set of elaborately designed convolutional neural networks (CNNs) and a three-layer stacked auto-encoder (SAE). The set of CNNs extracts complementary facial features from multimodal data. Then, the extracted features are concatenated to form a high-dimensional feature vector, whose dimension is compressed by SAE. All the CNNs are trained using a subset of 9,000 subjects from the publicly available CASIA-WebFace database, which ensures the reproducibility of this work. Using the proposed single CNN architecture and limited training data, 98.43% verification rate is achieved on the LFW database. Benefited from the complementary information contained in multimodal data, our small ensemble system achieves higher than 99.0% recognition rate on LFW using publicly available training set.

Index Terms—Face recognition, deep learning, convolutional neural networks, multimodal system.

I. INTRODUCTION

FACE recognition has been one of the most extensively studied topics in computer vision. The importance of face recognition is closely related to its great potential in multimedia applications, *e.g.*, photo album management in social networks, human machine interaction, and digital entertainment. With years of effort, significant progress has been achieved for face recognition. However, it remains a challenging task for multimedia applications, as observed in recent works [1], [2]. In this paper, we handle the face recognition problem for matching internet face images appeared in social networks, which is one of the most common applications in multimedia circumstances.

Recognizing the face images appeared in social networks is difficult, due to the reasons mainly from the following two perspectives. First, the face images uploaded to social networks are captured in real-world conditions; therefore faces in these images usually exhibit rich variations in pose, illumination, expression, and occlusion, as illustrated in Fig. 1. Second, face recognition in social networks is a large-scale recognition problem due to the numerous face images of potentially large amount of users. The prediction accuracy of face recognition

C. Ding and D. Tao are with the Centre for Quantum Computation and Intelligent Systems, and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 81 Broadway, Ultimo, NSW 2007, Australia (email: changxing.ding@student.uts.edu.au, dacheng.tao@uts.edu.au).

Fig. 1. Face images in multimedia applications usually exhibit rich variations in pose, illumination, expression, and occlusion.

algorithms usually degrades dramatically with the increase of face identities.

Accurate face recognition depends on high quality face representations. Good face representation should be discriminative to the change of face identify while remains robust to intra-personal variations. Conventional face representations are built on local descriptors, *e.g.*, Local Binary Patterns (LBP) [3], Local Phase Quantization (LPQ) [4], [5], Dual-Cross Patterns (DCP) [6], and Binarised Statistical Image Features (BSIF) [7]. However, the representation composed by local descriptors is too shallow to differentiate the complex nonlinear facial appearance variations. To handle this problem, recent works turn to Convolutional Neural Networks (CNNs) [8], [9] to automatically learn effective features that are robust to the nonlinear appearance variation of face images. However, the existing works of CNN on face recognition extract features from limited modalities, the complementary information contained in more modalities is not well studied.

Inspired by the complementary information contained in multi-modalities and the recent progress of deep learning on various fields of computer vision, we present a novel face representation framework that adopts an ensemble of CNNs to leverage the multimodal information. The performance of the proposed multimodal system is optimized from two perspectives. First, the architecture for single CNN is elaborately designed and optimized with extensive experimentations. Second, a set of CNNs is designed to extract complementary information from multiple modalities, *i.e.*, the holistic face image, the rendered frontal face image by 3D model, and uniformly sampled face patches. Besides, we design different structures for different modalities, *i.e.*, a complex structure is designed for the modality that contains the richest information while a simple structure is proposed for the modalities with

less information. In this way, we strike a balance between recognition performance and efficiency. The capacity of each modality for face recognition is also compared and discussed.

We term the proposed deep learning-based face representation scheme as Multimodal Deep Face Representation (MM-DFR), as illustrated in Fig. 2. Under this framework, the face representation of one face image involves feature extraction using each of the designed CNNs. The extracted features are concatenated as the raw feature vector, whose dimension is compressed by a three-layer SAE. Extensive experiments on the Labeled Face in the Wild (LFW) [10] and CASIA-WebFace databases [11] indicate that superior performance is achieved with the proposed MM-DFR framework. Besides, the influence of several implementation details, *e.g.*, the usage strategies of ReLU nonlinearity, multiple modalities, aggressive data augmentation, multi-stage training, and L2 normalization, is compared and discussed in the experimentation section. To the best of our knowledge, this is the first published approach that achieves higher than 99.0% recognition rate using a publicly available training set on the LFW database.

The remainder of the paper is organized as follows: Section II briefly reviews related works for face recognition and deep learning. The proposed MM-DFR face representation scheme is illustrated in Section III. Face matching using MM-DFR is described in Section IV. Experimental results are presented in Section V, leading to conclusions in Section VI.

II. RELATED STUDIES

A. Face Image Representation

Popular face representations can be broadly grouped into two categories: local descriptor-based representations and deep learning-based representations.

Traditional face representations are based on local descriptors [12], [13]. Local descriptors can be further divided into two groups: the handcrafted descriptors and the learning-based descriptors. Among the handcrafted descriptors, Ahonen *et al.* [3] proposed to employ the texture descriptor LBP for face representation. LBP works by encoding the gray-value difference between each pixel and its neighboring pixels into binary codes. Ding *et al.* [6] proposed the Dual-Cross Patterns (DCP) descriptor to encode second order statistics along the distribution directions of facial components. Other effective handcrafted local descriptors include Local Phase Quantization (LPQ) [4] and Gabor-based descriptors. Representative learning-based descriptors include Binarised Statistical Image Features (BSIF) [7], [14] and Discriminant Face Descriptor (DFD) [15], *et al.*. Compared with the handcrafted descriptors, the learning-based descriptors usually optimize the pattern encoding step using machine learning techniques. An extensive and systematic comparison among existing local descriptors for face recognition can be found in [6]; and a detailed summarization on local descriptor-based face representations can be found in a recent survey [1]. Despite of its ease of use, the local descriptor-based approaches have clear limitations: the constructed face reprsentation is sensitive to the non-linear intra-personal variations, *e.g.*, pose [16], expression [17], and illumination [13]. In particular, the intra-personal appearance

change caused by pose variations may substantially surpass the difference caused by identities [16].

The complicated facial appearance variations call for non-linear techniques for robust face representation, and recent progress on deep learning provides an effective tool. In the following, we review the most relevant progress for deep learning-based face recognition. Taigman *et al.* [8] proposed the DeepFace architecture for face recognition. They use the softmax loss, *i.e.*, the face identification loss, as the supervisory signal to train the network and achieve high recognition accuracy approaching the human-level. Sun *et al.* [9] proposed to combine the identification and verification losses for more effective training. They empirically verified that the combined supervisory signal is helpful to promote the discriminative power of extracted CNN features. Zhou *et al.* [18] investigated the influence of distribution and size of training data to the performance of CNN. With a huge training set composed of 5 millions of labelled faces, they achieved an accuracy of 99.5% accuracy on LFW using naive CNN structures. One common problem for the above works is that they all employ private face databases for training. Due to the distinct size and unknown distribution of these private data, the performance of the above works may not be directly comparable. Recently, Yi *et al.* [11] released the CASIA-WebFace database which contains 494,414 labeled images of 10,575 subjects. The availability of such a large-scale database enables researchers to compete on a fair starting line. In this paper, the training of all CNNs are conducted exclusively on a subset of 9,000 subjects of the CASIA-WebFace database, which ensures the reproducibility of this work. The CNN architectures designed in this paper are inspired by two previous works [19], [11], but with a number of modifications and improvements, and our designed CNN models have visible advantage in performance.

B. Multimodal-based Face Recognition

Most of face recognition algorithms extract a single face representation from the face image. However, they are restrictive in capturing the diverse information contained in the face image. To handle this problem, Ding *et al.* [6] proposed to extract the Multi-directional Multi-level DCPs (MDML-DCPs) feature which includes three holistic-level features and six component-level features. The set of the nine facial features composes the face representation. Similar strategies have been adopted in deep learning-based face representations. For example, the DeepFace approach [8] adopts the same CNN structure to extract facial features from RGB image, gray-level image and gradient map. The set of face representations are fused in the score level. Sun *et al.* [9] proposed to extract deep features from 25 image patches cropped with various scales and positions. The dimension of the concatenated deep features is reduced by Principle Component Analysis (PCA). Multimodal systems that fuse multiple feature cues are also employed in other topics of multimedia and computer vision, *e.g.*, visual tracking [20], image classification [21], [22], [23], and social media analysis [24], [25], [26], [27], [28].

Our multimodal face recognition system is related to the previous approaches, and there is clear novelty. First, we
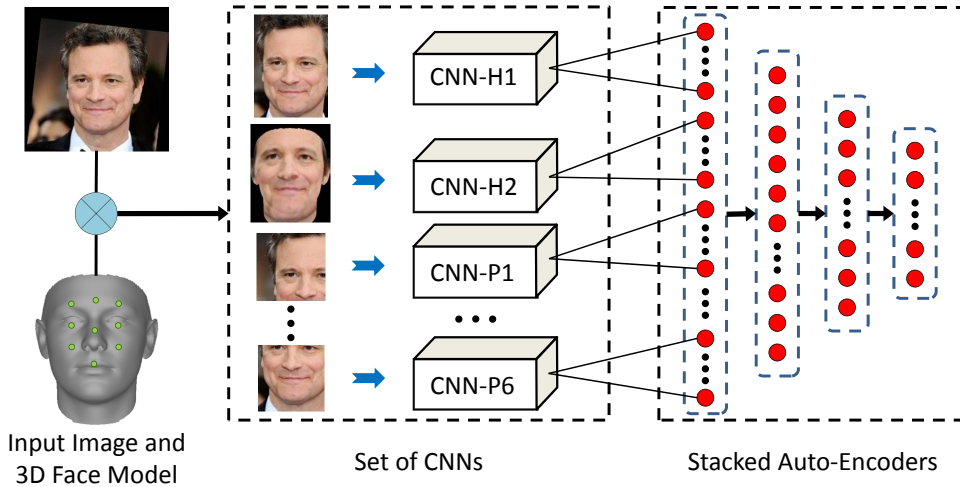
Fig. 2. Flowchart of the proposed multimodal deep face representation (MM-DFR) framework. MM-DFR is essentially composed of two steps: multimodal feature extraction using a set of CNNs, and feature-level fusion of the set of CNN features using SAE.

extract multimodal features from the holistic face image, rendered frontal face by 3D face model, and uniformly sampled image patches. The three modalities stand for holistic facial features and local facial features, respectively. Different from [8] that employs the 3D model to assist 2D piece-wise face warping, we utilize the 3D model to render a frontal face in 3D domain, which indicates much stronger alignment compared with [8]. Different from [9] that randomly crops 25 patches over the face image using dense facial feature points, we uniformly sample a small number of patches with the help of 3D model and sparse facial landmarks, which is more reliable compared with dense landmarks. Second, we propose to employ SAE to compress the high-dimensional deep feature into a compact face signature. Compared with the traditional PCA approach for dimension reduction, SAE has advantage in learning non-linear feature transformations. Third, the large-scale unconstrained face identification problem has not been well studied due to the lack of appropriate face databases. Fortunately, the recently published CASIA-WebFace [11] database provides the possibility for such kind of evaluation. In this paper, we evaluate the identification performance of MM-DFR on the CASIA-WebFace database.

III. MULTIMODAL DEEP FACE REPRESENTATION

In this section, we describe the proposed MM-DFR framework for face representation. As shown in Fig. 2, MM-DFR is essentially composed of two steps: multimodal feature extraction using a set of CNNs, and feature-level fusion of the set of CNN features using SAE. In the following, we describe the two main components in detail.

A. Single CNN Architecture

All face images employed in this paper are first normalized to 230×230 pixels with an affine transformation according to the coordinates of five sparse facial feature points, *i.e.*, both eye centers, the nose tip, and both mouth corners. Sample images after the affine transformation are illustrated in Fig. 1. We employ an off-the-shelf face alignment tool [29] for facial

feature detection. Based on the normalized image, one holistic face image of size 165×120 pixels (Fig. 3a) and six image patches of size 100×100 pixels (Fig. 3b) are sampled. Another holistic face image is obtained by 3D pose normalization using OpenGL [16]. Pose variation is reduced in the rendered frontal face, as shown in Fig. 3a.

Two CNN models named NN1 and NN2 are designed, which are closely related to the ones proposed in [19], [11], but with a number of modifications and improvements. We denote the CNN that extracts feature from the holistic face image as CNN-H1. In the following, we take CNN-H1 for example to illustrate the architectures of NN1 and NN2, as shown in Table I and Table II, respectively. The other seven CNNs employ similar structure but with modifications in resolution for each layer. The major difference between NN1 and NN2 is that NN2 is both deeper and wider than NN1. With larger structure, NN2 is more robust to highly non-linear facial appearance variations; therefore, we apply it to CNN-H1. NN1 is smaller but more efficient and we apply it to the other seven CNNs, with the underlying assumption that the image patches and pose normalized face contain less nonlinear appearance variations. Compared with NN1, NN2 is more vulnerable to overfitting due to its larger number of parameters. In this paper, we make use of aggressive data augmentation and multi-stage training strategies to reduce overfitting. Details of the two strategies are described in the experimentation section.

NN1 contains 10 convolutional layers, 4 max-pooling layers, 1 mean-pooling layer, and 2 fully-connected layers. In comparison, NN2 incorporates 12 convolutional layers. Small filters of 3×3 are utilized for all convolutional layers. As argued in [19], successive convolutions by small filters equal to one convolution operation by a large filter, but effectively enhances the model's discriminative power and reduces the number of filter parameters to learn. ReLU [30] activation function is utilized after all but the last convolutional layers. The removal of ReLU nonlinearity helps to generate dense features, as described in [11]. We also remove the ReLU nonlinearity after Fc6; therefore the projection of convolutional

TABLE I
DETAILS OF THE MODEL ARCHITECTURE FOR NN1

Name	Type	Input Size	Filter Number	Filter Size /stride /pad	With Relu
Conv11	conv	165×120	64	3×3 /1 /0	yes
Conv12	conv	163×118	128	3×3 /1 /0	yes
Pool1	max pool	161×116	N/A	2×2 /2 /0	no
Conv21	conv	80×58	64	3×3 /1 /0	yes
Conv22	conv	78×56	128	3×3 /1 /0	yes
Pool2	max pool	76×54	N/A	2×2 /2 /0	no
Conv31	conv	38×27	64	3×3 /1 /1	yes
Conv32	conv	38×27	128	3×3 /1 /1	yes
Pool3	max pool	38×27	N/A	2×2 /2 /1	no
Conv41	conv	20×14	128	3×3 /1 /1	yes
Conv42	conv	20×14	256	3×3 /1 /1	yes
Pool4	max pool	20×14	N/A	2×2 /2 /0	no
Conv51	conv	10×7	128	3×3 /1 /1	yes
Conv52	conv	10×7	256	3×3 /1 /1	**no**
Pool5	mean pool	10×7	N/A	2×2 /2 /1	no
Dropout	dropout	6144×1	N/A	N/A	N/A
Fc6	fully-conn	512×1	N/A	N/A	**no**
Fc7	fully-conn	9000×1	N/A	N/A	no
Softmax	softmax	9000×1	N/A	N/A	N/A

TABLE II
DETAILS OF THE MODEL ARCHITECTURE FOR NN2

Name	Type	Input Size	Filter Number	Filter Size /stride /pad	With Relu
Conv11	conv	165×120	64	3×3 /1 /0	yes
Conv12	conv	163×118	128	3×3 /1 /0	yes
Pool1	max pool	161×116	N/A	2×2 /2 /0	no
Conv21	conv	80×58	64	3×3 /1 /0	yes
Conv22	conv	78×56	128	3×3 /1 /0	yes
Pool2	max pool	76×54	N/A	2×2 /2 /0	no
Conv31	conv	38×27	128	3×3 /1 /1	yes
Conv32	conv	38×27	128	3×3 /1 /1	yes
Pool3	max pool	38×27	N/A	2×2 /2 /1	no
Conv41	conv	20×14	256	3×3 /1 /1	yes
Conv42	conv	20×14	256	3×3 /1 /1	yes
Conv43	conv	20×14	256	3×3 /1 /1	yes
Pool4	max pool	20×14	N/A	2×2 /2 /0	no
Conv51	conv	10×7	256	3×3 /1 /1	yes
Conv52	conv	10×7	256	3×3 /1 /1	yes
Conv53	conv	10×7	256	3×3 /1 /1	**no**
Pool5	mean pool	10×7	N/A	2×2 /2 /1	no
Dropout	dropout	6144×1	N/A	N/A	N/A
Fc6	fully-conn	512×1	N/A	N/A	**no**
Fc7	fully-conn	9000×1	N/A	N/A	no
Softmax	softmax	9000×1	N/A	N/A	N/A

features by Fc6 layer is from dense to dense, which means that Fc6 effectively equals to a linear dimension reduction layer that is similar to PCA or Linear Discriminative Analysis (LDA). This is different from previous works that favor sparse features produced by ReLU [8], [9], [31]. Our model is also different from [11] since [11] simply removes the linear dimension reduction layer (Fc6). The output of the Fc6 layer is employed as face representation. In the experimental section, we empirically justify that the dense-to-dense projection by Fc6 is advantageous to produce more discriminative features. The forward function of ReLU is represented as

$$R(x_i) = max(0, W_c^T x_i + b_c), \qquad (1)$$

where x_i, W_c, and b_c are the input, weight, and bias of the corresponding convolutional layer before the ReLU activation function. $R(x_i)$ is the output of the ReLU activation function. The dimension of the Fc6 layer is set to 512. The dimension of the Fc7 is set to 9000, which equals to the number of training subjects employed in this paper. We employ dropout [32] as a regularizer on the first fully-connected layer in the case of overfitting caused by the large amount of parameters. The dropout ratio is set to 0.4. Since this low-dimensional face representation is utilized to distinguish as large as 9,000 subjects in the training set, it should be very discriminative and has good generalization ability.

The other holistic image is rendered by OpenGL with the help of 3D generic face model [16]. Pose variation is reduced in the rendered image. We denote the CNN that extracts deep feature from this image as CNN-H2, as illustrated in Fig. 2. Therefore, the first two CNNs encode holistic image features from different modalities. The CNNs that extract features from the six image patches are denoted as CNN-P1, CNN-P2, to
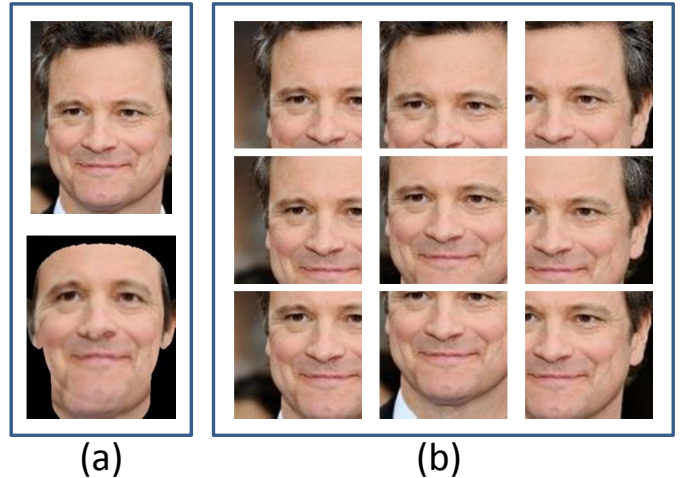

Fig. 3. The normalized holistic face images and image patches as input for MM-DFR. (a) The original holistic face image and the 3D pose normalized holistic image; (b) Image patches uniformly sampled from the original face image. Due to facial symmetry and the augmentation by horizontal flipping, we only leverage the six patches illustrated in the first two columns.

CNN-P6, respectively, as illustrated in Fig. 2. Exactly the same network structure is adopted for each of the six CNNs.

Different from previous works that randomly sample a large number of image patches [9], we propose to sample a small number of image patches uniformly in the semantic meaning, which contributes to maximizing the complementary information contained within the sampled patches. However, the uniform sampling of the image patches is not easy due to the pose variations of the face appeared in real-world images,
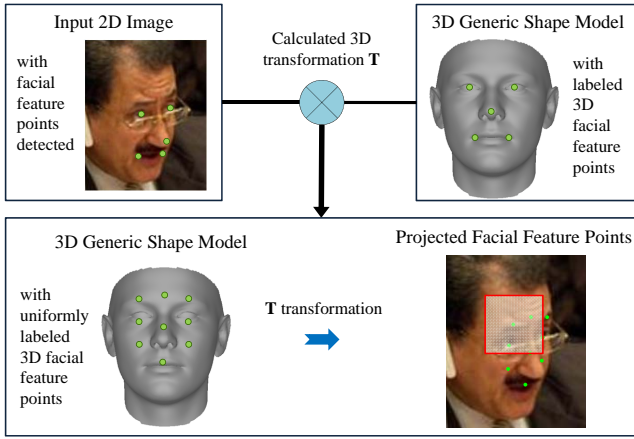
Fig. 4. The principle of patch sampling adopted in this paper. A set of 3D landmarks are uniformly labeled on the 3D face model, and are projected to the 2D image. Centering around each landmark, a square patch of size 100×100 pixels is cropped, as illustrated in Fig. 3b.

Fig. 5. More examples about the uniformly detected landmarks that are projected from a generic 3D face model to 2D images.

as shown in Fig. 1. We tackle this problem with a recently proposed strategy for pose-invariant face recognition [33]. The principle of the patch sampling process is illustrated in Fig. 4. In brief, nine 3D landmarks are manually labeled on a generic 3D face model and the 3D landmarks spread uniformly across the face model. In this paper, we consistently employ the mean shape of the Basel Face Model as the generic 3D face model [34]. Given a 2D face image, it is first aligned to the generic 3D face model using orthogonal projection with the help of five facial feature points. Then, the pre-labeled 3D landmarks are projected to the 2D image. Lastly, a patch of size 100×100 pixels is cropped centering around each of the projected 2D landmarks. More examples of the detected 2D uniform landmarks are shown in Fig. 5. It is clear that the patches are indeed uniformly sampled in the semantic meaning regardless of the pose variations of the face image.

B. Combination of CNNs using Stacked Auto-Encoder

We denote the features extracted by the set of CNNs as $\{x_1, x_2, \cdots, x_K\}$, where $x_i \in \mathbb{R}^{d \times 1}, 1 \leq i \leq K$. In this paper, K equals to 8 and d equals to 512. The set of features represents multimodal information for face recognition. We conduct feature-level fusion to obtain a single signature for each face image. In detail, the features extracted by the eight

CNNs are concatenated as a large feature vector, denoted as:

$$\hat{x} = [x_1; x_2; \cdots; x_K] \in \mathbb{R}^{Kd \times 1}. \tag{2}$$

$\hat{x}$ is high dimensional, which is impractical for real-world face recognition applications. We further propose to reduce the dimension of $\hat{x}$ by SAE. Compared with the traditional dimension reduction approaches, $e.g.$, PCA, SAE has advantage in learning non-linear feature transformations. In this paper, we employ a three-layer SAE. The number of the neurons of the three auto-encoders are 2048, 1024, and 512, respectively. The output of the last encoder is utilized as the compact signature of the face image. The structure for the designed SAE is illustrated in Fig. 2.

Nonlinear activation function is utilized after each of the fully-connected layers. Two activation functions, $i.e.$, sigmoid function and hyperbolic tangent (tanh) function, are evaluated. The forward function of the sigmoid activation function is represented as

$$S(x_i) = \frac{1}{1 + exp(-W_f^T x_i - b_f)}. \tag{3}$$

The forward function of the tanh activation function is represented as

$$T(x_i) = \frac{exp(W_f^T x_i + b_f) - exp(-W_f^T x_i - b_f)}{exp(W_f^T x_i + b_f) + exp(-W_f^T x_i - b_f)}, \tag{4}$$

where x_i, W_f, and b_f are the input, weight, and bias of the corresponding fully-connected layer before the activation function. Different normalization schemes of $\hat{x}$ are adopted for the sigmoid and tanh activation functions, since their output space is different. For the sigmoid function, we normalize the elements of $\hat{x}$ to be within $[0, 1]$. For the tanh function, we normalize the elements of $\hat{x}$ to be within $[-1, +1]$. In the experimentation section, we empirically compare the performance of SAE with the two different nonlinearities.

IV. FACE MATCHING WITH MM-DFR

In this section, the face matching problem is addressed based on the proposed MM-DFR framework. Two evaluation modes are adopted: the unsupervised mode and the supervised mode. Suppose two features produced by MM-DFR for two images are denoted as y_1 and y_2, respectively. In the unsupervised mode, the cosine distance is employed to measure the similarity s between y_1 and y_2.

$$s(y_1, y_2) = \frac{y_1^T y_2}{\|y_1\| \|y_2\|}. \tag{5}$$

For the supervised mode, a number of discriminative or generative models can be employed [35], [36], [37], In this paper, we employ the Joint Bayesian (JB) model [36] as it is shown to outperform other popular models in recent works [6]. For both the unsupervised and supervised modes, the nearest neighbor (NN) classifier is adopted for face identification. JB models the face generation process as

$$x = \mu + \varepsilon, \tag{6}$$

where μ represents the identity of the subject, while ε represents intra-personal noise.

JB solves the face identification or verification problems by computing the log-likelihood ratio between the probability $P(x_1, x_2|H_I)$ that two faces belong to the same subject and the probability $P(x_1, x_2|H_E)$ that two faces belong to different subjects.

$$r(x_1, x_2) = log \frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)}, \tag{7}$$

where $r(x_1, x_2)$ represents the log-likelihood ratio, and we refer to $r(x_1, x_2)$ as similarity score for clarity in the experimental part of the paper.

V. EXPERIMENTAL EVALUATION

In this section, extensive experiments are conducted to present the effectiveness of the proposed MM-DFR framework. The experiments are conducted on two large-scale unconstrained face databases, *i.e.*, LFW [10] and CASIA-WebFace [11]. Images in both databases are collected from internet; therefore they are real images that appear in multimedia circumstances.

The LFW [10] database contains 13,233 images of 5,749 subjects. Images in this database exhibit rich intra-personal variations of pose, illumination, and expression. It has been extensively studied for the research of unconstrained face recognition in recent years. Images in LFW are organized into two "Views". View 1 is for model selection and parameter tuning while View 2 is for performance reporting. In this paper, we follow the official protocol of LFW and report the mean verification accuracy and the standard error of the mean (S_E) by the 10-fold cross-validation scheme on the View 2 data.

Despite of its popularity, the LFW database contains limited number of images and subjects, which restricts its evaluation for large-scale unconstrained face recognition applications. The CASIA-WebFace [11] database has been released recently. CASIA-WebFace contains 494,414 images of 10,575 subjects. As images in this database are collected in a semi-automatic way, there is a small amount of mis-labeled images in this database. Because there is no officially defined protocol for face recognition on this database, we define our own protocol for face identification in this paper. In brief, we divide CASIA-WebFace into two sets: a training set and a testing set. The 10,575 subjects are ranked in the descent order by the number of their images contained in the database. The 471,592 images of the top 9,000 subjects compose the training set. The 22,822 images of the rest 1,575 subjects make up the testing set.

All CNNs and SAE in this paper are trained using the 9,000 subjects in the defined training set above. Images are converted to gray-scale and geometrically normalized as described in Section III. For NN1, we double the size of the training set by flipping all training images horizontally to reduce overfitting. Therefore, the size of training data for NN1 is 943,184. For NN2, we adopt much more aggressive data augmentation by horizontal flipping, image jittering[1], and image downsampling. The size of the augmented training data for NN2

[1] For image jittering, we add random gaussian noise on the coordinates of the five facial feature points. The noise is distributed with zero mean and standard deviation of four pixels.
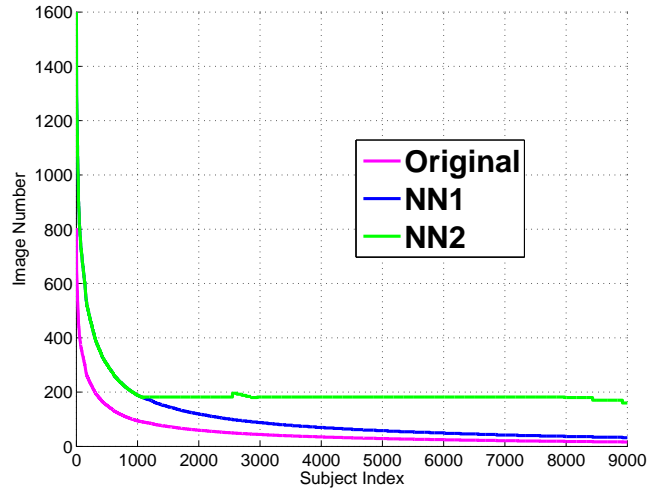

Fig. 6. Training data distribution for NN1 and NN2. This figure plots the number of images for each subject in the training set. The long-tail distribution characteristic [18] of the original training data is improved after the aggressive data augmentation for NN2.

is about 1.8 million. The distribution of training data for NN1 and NN2 is illustrated in Fig. 6. It is shown that the long-tail distribution characteristic [18] of the original training data is improved after the aggressive data augmentation for NN2.

We adopt the following multi-stage training strategy to train all the CNN models. First, we train the CNN models as a multi-class classification problem, *i.e.*, softmax loss is employed. For all CNNs, the initial learning rate for all learning layers is set to 0.01, and is divided by 10 after 10 epochs, to the final rate of 0.001. Second, we adopt the recently proposed triplet loss [38] for fine-tuning for 2 more epochs. We set the margin for the triplet loss to be 0.2 and learning rate to be 0.001. It is expected that this multi-stage training strategy can boost performance while converge faster than using the triplet loss alone [38]. For SAE, the learning rate decreases from 0.01 to 0.00001, gradually. We train each of the three auto-encoders one by one and each auto-encoder is trained for 10 epochs. In the testing phase, we extract deep feature from both the original image and its horizontally flipped image. Unless otherwise specified, the two feature vectors are averaged as the representation of the input face image. The open-source deep learning toolkit Caffe [39] is utilized to train all the deep models.

Five sets of experiments are conducted. First, we empirically justify the advantage of dense features for face recognition by excluding two ReLU nonlinearities compared with previous works. The performance of the proposed single CNN model is also compared against the state-of-the-art CNN models on the LFW database. Next, the performance of the eight CNNs contained within the MM-DFR framework is compared on face verification task on LFW. Then, the fusion of the eight CNNs by SAE is conducted and different nonlinearities are also compared. We also test the performance of MM-DFR followed with the supervised classifier JB. Lastly, face identification experiment is conducted on the CASIA-WebFace database with our own defined evaluation protocol.
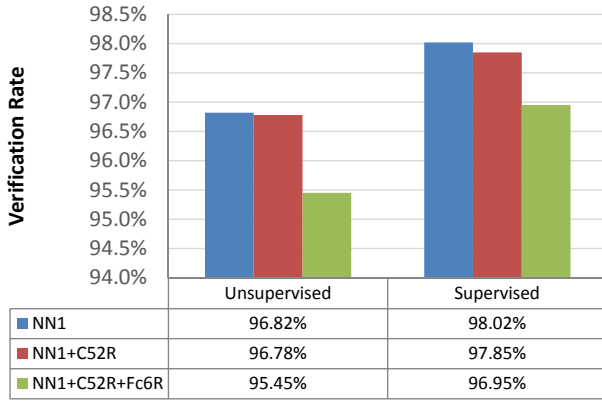
Fig. 7. Performance comparison on LFW with different usage strategies of ReLU nonlinearity.

	Accuracy (Unsupervised)	Accuracy (Supervised)
DeepFace [8]	95.92 ± 0.29	97.00 ± 0.87
DeepID2 [9]	-	$96.33 \pm -$
Arxiv2014 [11]	96.13 ± 0.30	97.73 ± 0.31
Facebook [41]	-	$98.00 \pm -$
MSU TR [40]	96.95 ± 1.02	97.45 ± 0.99
Ours (NN1)	$\mathbf{97.32 \pm 0.34}$	$\mathbf{98.05 \pm 0.22}$
Ours (NN2)	$\mathbf{98.12 \pm 0.24}$	$\mathbf{98.43 \pm 0.20}$

Fig. 8. ROC curves of different usage strategies of the ReLU nonlinearity on LFW.

A. Performance Comparison with Single CNN Model

In this experiment, we evaluate the role of ReLU nonlinearity using CNN-H1 as an example. For fast evaluation, the comparison is conducted with the simple NN1 structure described in Table I and only the softmax loss is employed for model training. Performance of CNN-H1 using the NN2 structure can be found in Table IV. Two paradigms[2] are followed: 1) the unsupervised paradigm that directly calculate the similarity between two CNN features using cosine distance metric. 2) the supervised paradigm that uses JB to calculate the similarity between two CNN features. For the supervised paradigm, we concatenate the CNN features of the original face image and its horizontally flipped version as the raw representation of each test sample. Then, we adopt PCA for dimension reduction and JB for similarity calculation. The dimension of the PCA subspace is tuned on the View 1 data of LFW and applied to the View 2 data. Both PCA and JB are trained on the CASIA-WebFace database. For PCA, to boost performance, we also re-evaluate the mean of CNN features using the 9 training folds of LFW in 10-fold cross validation.

The performance of three structures are reported in Fig. 7 and Fig. 8: 1) NN1, 2) NN1 with ReLU after Conv52 layer (denoted as NN1+C52R), and 3) NN1 with ReLU after both Conv52 and Fc6 (denoted as NN1+C52R+Fc6R). For both NN1+C52R and NN1+C52R+Fc6R, we replace the average pooling layer after Conv 52 with max pooling accordingly. It is shown in Fig. 7 that the ReLU nonlinearity after Conv52 or Fc6 actually harms the performance of CNN. The experimental results have two implications: 1) dense feature is preferable than sparse feature for CNN, as intuitively advocated in [11]. However, there is no experimental justification in [11]. 2) the linear projection from the output of the ultimate convolutional layer (Conv52) to the low-dimensional subspace (Fc6) is better than the commonly adopted non-linear projection. This is clear evidence that the negative response of the ultimate convolutional layer (Conv52) also contains useful information.

The performance by single CNN models on LFW is reported in Table. III. The performance of the state-of-the-art CNN

models is also tabulated. Compared with Fig. 7, we further improve the performance of NN1 by fine-tuning with triplet loss. It seems that the triplet loss mainly improves the performance for the unsupervised mode in our experiment. It is shown that the proposed CNN model consistently outperforms the state-of-the-art CNN models under both the unsupervised paradigm and supervised paradigm. In particular, compared with [11], [40] that all employ the complete CASIA-WebFace database for CNN training, we only leverage a subset of the CASIA-WebFace database. With more training data, we expect the proposed CNN model can outperform the other models with an even larger margin.

B. Performance of the Eight CNNs in MM-DFR

In this experiment, we present in Table IV the performance achieved by each of the eight CNNs contained within the MM-DFR framework. We report the performance of CNN-H1 with the NN2 structure while the other seven CNNs all employ the more efficient NN1 structure. The same as the previous experiment, both the unsupervised paradigm and supervised paradigm are followed. For the supervised paradigm, the PCA subspace dimension of the eight CNNs is unified to be 110. Besides, features of the original face image and the horizontally flipped version are L2 normalized before concatenation.

[2]Similar to previous works [8], [11], both the two paradigms defined in this paper correspond to the "Unrestricted, Labeled Outside Data Results" protocol that is officially defined in [10].

TABLE IV
PERFORMANCE COMPARISON ON LFW OF EIGHT INDIVIDUAL CNNs

	Accuracy (Unsupervised)	Accuracy (Supervised)
CNN-H1	**98.12 ± 0.24**	**98.43 ± 0.20**
CNN-H2	96.47 ± 0.44	97.67 ± 0.28
CNN-P1	96.83 ± 0.26	97.30 ± 0.22
CNN-P2	97.25 ± 0.31	98.00 ± 0.24
CNN-P3	96.70 ± 0.25	97.82 ± 0.16
CNN-P4	96.17 ± 0.31	96.93 ± 0.21
CNN-P5	96.05 ± 0.27	97.23 ± 0.20
CNN-P6	95.58 ± 0.17	96.72 ± 0.21

We find that this normalization operation typically boosts the performance of the supervised paradigm by 0.1% to 0.4%.

When combining Table III and Table IV, it is clear that CNN-H1 outperforms CNN-H2 with the same NN1 structure, although they both extract features from holistic face images. This maybe counter-intuitive, since the impact of pose variation has been reduced for CNN-H2. We explain this phenomenon from the following two aspects: 1) most images in LFW are near-frontal face images, so the 3D pose normalization employed by CNN-H2 does not contribute much to pose correction. 2) the errors in pose normalization bring about undesirable distortions and artifacts to facial texture, *e.g.*, the distorted eyes, nose, and mouth shown in Fig. 3(a). The distorted facial texture is adverse to face recognition, as argued in our previous work [1]. However, we empirically observe that the performance of MM-DFR drops slightly on View 1 data if we exclude CNN-H2, which indicates CNN-H2 provides complementary information to CNN-H1 from a novel modality. The contribution of CNN-H2 to MM-DFR is also justified by the last experiment in this section. Besides, the performance of the patch-level CNNs, *i.e.*, CNN-P1 to CNN-P6, fluctuates according to the discriminative power of the corresponding patches.

C. Fusion of CNNs with SAE

In this experiment, we empirically choose the best non-linearity for SAE that is employed for feature-level fusion of the eight CNNs. The structure of SAE employed in this paper is described in Fig. 2. For each CNN, we average the features of the original image and the horizontally flipped version. L2 normalization is conducted for each averaged feature before concatenating the features produced by the eight CNNs. Similar to the previous experiment, we find this normalization operation promotes the performance of MM-DFR. The dimension of the input for SAE is 4,096. Two types of non-linearities are evaluated, the sigmoid non-linearity and the tanh non-linearity, denoted as SAE-SIG and SAE-TANH, respectively. The output of the third encoder (before the nonlinear layer) is utilized as the signature of the face image. Cosine distance is employed to evaluate the similarity between two face images. SAE are trained on the training set of CASIA-WebFace, using feature vectors extracted from both the original images and the horizontally flipped ones. The performance of SAE-SIG and SAE-TANH is 98.33% and
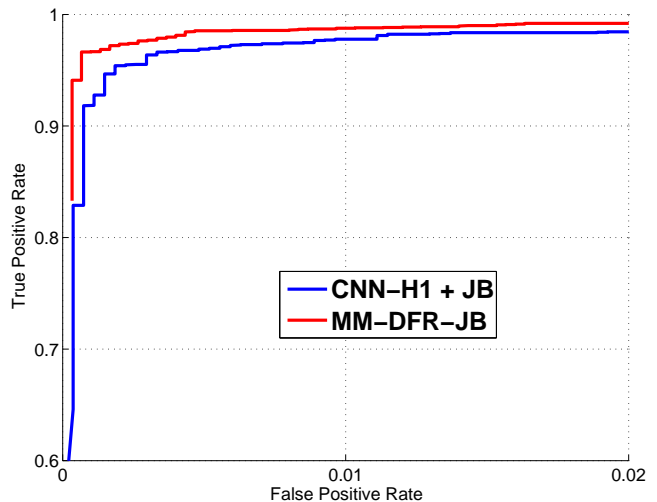

Fig. 9. Performance comparison between the proposed MM-DFR approach and single modality-based CNN on the face verification task.

97.90% on the View1 data of LFW, respectively.

SAE-TANH considerably outperforms SAE-SIG. One important difference between the sigmoid non-linearity and the tanh non-linearity is that they normalize the elements of the feature to be within $[0, 1]$ and $[-1, 1]$, respectively. Compared with the tanh non-linearity, the sigmoid non-linearity loses the sign information of feature elements. However, the sign information is valuable for discriminative power.

D. Performance of MM-DFR with Joint Bayesian

The above three experiments have justified the advantage of the proposed CNN structures. In this experiment, we further promote the performance of the proposed framework.

We show the performance of MM-DFR with JB, where the output of MM-DFR is utilized as the signature of the face image. We term this face recognition pipeline as MM-DFR-JB. For comparison, the performance achieved by CNN-H1 with the JB classifier is also presented, denoted as "CNN-H1 + JB". The performance of the two systems is tabulated in Table V and the ROC curves are illustrated in Fig. 9. It is shown that MM-DFR considerably outperforms the single modal-based approach, which indicates the fusion of multimodal information is important to promote the performance of face recognition systems. By excluding the five labeling errors in LFW, the actual performance of MM-DFR-JB reaches 99.10%.

Our simple 8-net based ensemble system also outperforms DeepID2 [9], which includes as much as 25 CNNs. Some more recent approaches that were published after the submission of this paper, *e.g.* [38], [31], achieve better performance than MM-DFR. However, they either employ significantly larger private training dataset or considerably larger number of CNN models. In comparison, we employ only 8 nets and train the models using a relatively small training set.

E. Face Identification on CASIA-WebFace Database

The face identification experiment is conducted on the test data of the CASIA-WebFace database, which includes 22,822

TABLE V
PERFORMANCE EVALUATION OF MM-DFR WITH JB

	#Nets	Accuracy(%)$\pm S_E$
CNN-H1 + JB	1	98.43 ± 0.20
DeepFace [8]	7	97.25 ± 0.81
MSU TR [40]	7	98.23 ± 0.68
DeepID2 [9]	25	98.97 ± 0.25
MM-DFR-JB	8	**99.02 ± 0.19**

TABLE VI
THE RANK-1 IDENTIFICATION RATES BY DIFFERENT COMBINATIONS OF
MODALITIES ON CASIA-WEBFACE DATABASE

	Identification Rates
CNN-H1 + JB	72.26%
CNN-H2 + JB	69.07%
CNN-H1&H2 + JB	74.51%
CNN-P1 to P6 + JB	76.01%
MM-DFR-JB	76.53%

Fig. 10. CMS curves by different combinations of modalities on the face identification task.

images of 1,575 subjects. For each subject, the first five images are selected to make up the gallery set, which can generally be satisfied in many multimedia applications, *e.g.*, social networks where each subject has multiple face images. All the other images compose the probe set. Therefore, there are 7,875 gallery images and 14,947 probe images in total.

The rank-1 identification rates by different combinations of modalities are tabulated in Table VI. The corresponding Cumulative Match Score (CMS) curves are illustrated in Fig. 10. It is shown that although very high face verification rate has been achieved on the LFW database, large-scale face identification in real-world applications is still a very hard problem. In particular, the rank-1 identification rate by the proposed approach is only 76.53%.

It is clear that the proposed multimodal face recognition algorithm significantly outperforms the single modal based approach. In particular, the rank-1 identification rate of MM-DFR-JB is higher than that of "CNN-H1 + JB" by as much as 4.27%. "CNN-H1 + JB" outperforms "CNN-H2 + JB" with a large margin, partially because CNN-H1 is based on the larger architecture NN2 and trained with more aggressively augmented data. However, the combination of the two modalities still considerably boosts the performance by 2.25% on the basis of CNN-H1, which forcefully justifies the contribution of the new modality introduced by 3D pose normalization. These experimental results are consistent with those obversed on the LFW database. Experimental results on both datasets strongly justify the effectiveness of the proposed MM-DFR framework for multimedia applications.

VI. CONCLUSION

Face recognition in multimedia applications is a challenging task because of the rich appearance change caused by pose, expression, and illumination variations. We handle this problem by elaborately designing a deep architecture that employs complementary information from multimodal image data. First, we enhance the recognition ability of each CNN
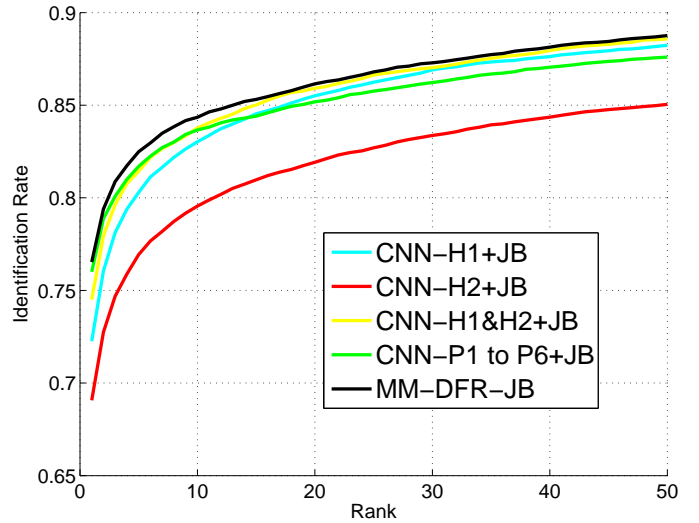
by carefully integrating a number of published or our own developed tricks, such as deep structures, small filters, careful use of ReLU nonlinearity, aggressive data augmentation, dropout, and multi-stage training with multiple losses, L2 normalization. Second, we propose to extract multimodal information using a set of CNNs from the original holistic face image, the rendered frontal pose image by 3D model, and uniformly sampled image patches. Third, we present the feature-level fusion approach using stacked auto-encoders to fuse the features extracted by the set of CNNs, which is advantageous to learn non-linear dimension reduction. Extensive experiments have been conducted for both face verification and face identification experiments. As the proposed MM-DFR approach effectively employs multimodal information for face recognition, clear advantage of MM-DFR is shown compared with the single modal-based algorithms and some state-of-the-art deep models. Other deep learning based approaches may also benefit from the structures that have been proved to be useful in this paper. In the future, we will try to integrate more multimodal information into the MM-DFR framework and further promote the performance of single deep architecture such as NN2.

REFERENCES

[1] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *arXiv preprint arXiv:1502.04383*, 2015.
[2] J. Y. Choi, W. De Neve, K. N. Plataniotis, and Y. M. Ro, "Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 14–28, 2011.
[3] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, 2006.

[4] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image Signal Process.* Springer, 2008, pp. 236–243.

[5] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," in *Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[6] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell., doi: 10.1109/TPAMI.2015.2462338*, 2015.

[7] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 1363–1366.

[8] 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.

[9] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.

[10] 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," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

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

[12] M. Gunther, A. Costa-Pazo, C. Ding *et al.*, "The 2013 face recognition evaluation in mobile environment," in *Proc. IEEE/IAPR Int. Conf. Biometrics*, 2013, pp. 1–7.

[13] R. Beveridge, H. Zhang, B. Draper *et al.*, "Report on the fg 2015 video person recognition evaluation," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognit.*, 2015.

[14] S. Rahimzadeh Arashloo and J. Kittler, "Dynamic texture recognition using multiscale binarised statistical image features," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2099–2109, 2014.

[15] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 289–302, Feb. 2014.

[16] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 980–993, 2015.

[17] C.-K. Hsieh, S.-H. Lai, and Y.-C. Chen, "Expression-invariant face recognition with constrained optical flow warping," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 600–610, 2009.

[18] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of lfw benchmark or not?" *arXiv preprint arXiv:1501.04690*, 2015.

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

[20] X. Mei, Z. Hong, D. Prokhorov, and D. Tao, "Robust multitask multiview tracking in videos," *IEEE Trans. Neural Netw. Learn. Syst.*, 2015.

[21] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Two-dimensional multilabel active learning with an efficient online adaptation model for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1880–1897, 2009.

[22] C. Xu, D. Tao, and C. Xu, "Large-margin multi-viewinformation bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, 2014.

[23] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Two-dimensional active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[24] G.-J. Qi, X.-S. Hua, and H.-J. Zhang, "Learning semantic distance from community-tagged media collection," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 243–252.

[25] E. Eaton, S. Jacob *et al.*, "Multi-view constrained clustering with an incomplete mapping between views," *Knowledge and information systems*, vol. 38, no. 1, pp. 231–257, 2014.

[26] G.-J. Qi, C. Aggarwal, Q. Tian, H. Ji, and T. S. Huang, "Exploring context and content links in social media: A latent space method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 850–862, 2012.

[27] S. F. Hussain and S. Bashir, "Co-clustering of multi-view datasets," *Knowledge and Information Systems*, pp. 1–26, 2015.

[28] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 17–26.

[29] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[30] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *Proc.*

[31] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2892–2900.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[33] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3539–3545.

[34] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.

[35] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[36] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 566–579.

[37] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell., doi: 10.1109/TPAMI.2015.2456899*, 2015.

[38] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[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 *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[40] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," *arXiv preprint arXiv:1507.07242*, 2015.

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Web-scale training for face identification," *arXiv preprint arXiv:1406.5266*, 2014.