Deep Learning for Biometrics: A Survey

KALAIVANI SUNDARARAJAN and DAMON L. WOODARD, University of Florida

In the recent past, deep learning methods have demonstrated remarkable success for supervised learning tasks in multiple domains including computer vision, natural language processing, and speech processing. In this article, we investigate the impact of deep learning in the field of biometrics, given its success in other domains. Since biometrics deals with identifying people by using their characteristics, it primarily involves supervised learning and can leverage the success of deep learning in other related domains. In this article, we survey 100 different approaches that explore deep learning for recognizing individuals using various biometric modalities. We find that most deep learning research in biometrics has been focused on face and speaker recognition. Based on inferences from these approaches, we discuss how deep learning methods can benefit the field of biometrics and the potential gaps that deep learning approaches need to address for real-world biometric applications.

CCS Concepts: • **Security and privacy → Biometrics**; • **Computing methodologies → Neural networks**; *Supervised learning by classification*;

Additional Key Words and Phrases: Deep learning, face recognition, speaker recognition, feature learning, convolutional neural networks, deep belief nets, autoencoders

ACM Reference format:
Kalaivani Sundararajan and Damon L. Woodard. 2018. Deep Learning for Biometrics: A Survey. *ACM Comput. Surv.* 51, 3, Article 65 (May 2018), 34 pages.
https://doi.org/10.1145/3190618

1 INTRODUCTION

Deep learning involves stacking multiple layers of learning algorithms to approximate highly nonlinear functions. This enables deep learning algorithms to learn hierarchical representations/features from data. This feature learning has largely replaced hand-engineered features in various domains including vision, speech, and natural language processing.

Deep learning owes its resurgence to efficient optimization techniques and powerful computational resources. Despite its huge success, deep learning has met with reluctant acceptance by researchers due to its scarce theoretical backing. However, industries with large amounts of data and computing resources have enthusiastically adopted deep learning. Now that deep learning has proved to be highly successful for supervised learning, researchers are exploring its capabilities in unsupervised learning, thereby taking a small step toward machine understanding. In this article, we survey the impact of deep learning in the field of biometrics and discuss the associated challenges.

Authors' addresses: K. Sundararajan, Department of Computer and Information Science and Engineering, University of Florida, 601 Gale Lemerand Dr., P.O.Box 116550, Gainesville, FL 32611; email: kalaivani.s@ufl.edu; D. L. Woodard, Department of Electrical and Computer Engineering, University of Florida, 601 Gale Lemerand Dr., P.O.Box 116550, Gainesville, FL 32611; email: dwoodard@ece.ufl.edu.

https://doi.org/10.1145/3190618

Biometrics deals with identifying people using their physical or behavioral characteristics. Physiological biometrics depend on the physical characteristics of a person (e.g., face, fingerprints, and iris). Behavioral biometrics depend on a person's activities and is hugely influenced by other socioenvironmental factors. For biometric systems to be effective in real-world applications, they need to overcome certain associated challenges:

- **Large number of identities:** Biometric recognition involves distinguishing potentially millions of individuals in whom distinguishing factors may be subtle. This requires highly complex models to identify individuals at such a scale.
- **Intraperson variations:** Typically, multiple samples of an individual are required to completely capture variations in characteristics. These *intraperson* variations may sometimes be larger than variations between individuals (i.e. *interperson* variations).
- **Noisy and distorted input:** Biometric data collected in real-world applications are quite noisy and distorted due to noisy biometric sensors or other factors. Data quality worsens when biometric data are captured covertly due to completely unconstrained conditions.
- **Extracting biometric information:** Extracting relevant biometric information from noisy input data requires significant preprocessing (e.g., extracting a person's face in a cluttered background or a person's speech signal from a noisy background).
- **Permanence:** Since biometric traits are based on human characteristics, they might not be consistent. Physiological biometrics vary gradually over time, while behavioral biometrics are additionally influenced by socio-environmental factors.
- **Uniqueness:** It is unclear whether a single biometric trait can uniquely identify a person. Specifically, most behavioral biometrics are not effective for identification and are being solely used for verification purposes.
- **Attacks on biometric systems:** Biometric systems are prone to attacks on various levels. It is imperative that such attacks be identified as part of the recognition system.

1.1 Motivation

Given the recent success of deep learning approaches, we hope that they might be beneficial in addressing some of the above-mentioned challenges in biometric recognition. Deep learning approaches learn features from the data, and, when trained discriminatively, they might learn subtle features that can distinguish between large numbers of individuals. Furthermore, if there are sufficient numbers of samples representative of different factors that impact recognition, deep learning techniques can learn to disentangle such factors while learning discriminative feature representations. This might help handle large intraclass variations and noisy biometric data. However, one of the major drawbacks is that, in order to capture all these variations, the model has to be sufficiently complex and hence requires large amounts of training data.

Huge efforts would be required to collect data that exhibit gradual variation over time (e.g., face datasets for age estimation). Under such scenarios, generative deep learning approaches may be used to synthesize such variations. Owing to its capability to learn from data, deep learning might also be useful in segmenting biometric data from noisy background. Given these possibilities, we investigate how biometric recognition approaches have taken advantage of the merits of deep learning and aspects where deep learning can help improve biometric systems.

2 DEEP LEARNING

2.1 Overview

Deep learning [9, 127] learns high-level abstractions in the data by stacking multiple learning layers. These layers are typically neural networks because of their ability to model highly nonlinear

functions. Recently, deep learning has outperformed previous state-of-the-art methods in various domains. The following driving factors are behind this success [76]:

- **Feature learning:** Deep learning methods learn features from data which help to generalize for other related tasks. Various correlated factors are disentangled in these learned features compared to hand-engineered features which are designed to be invariant to such factors.
- **Hierarchical representations:** These methods learn hierarchical representations: Lower level layers learn simple features, and higher level layers learn increasingly complex features composed of lower level features. This helps encode both local and global properties in the final feature representation.
- **Distributed representations:** The learned representations are distributed because a single factor can be explained by many neurons, and a single neuron can help explain many factors. This many-to-many relationship yields compact, dense representations that can generalize nonlocally thereby helping combat the curse of dimensionality.
- **Computational resources:** GPUs and other parallel computing resources have made possible the training of large, deep neural networks with millions of training examples. This has helped demonstrate the impact of deep neural networks in various domains.
- **Large-scale datasets:** Large-scale datasets with huge amounts of training samples have helped deep learning create a significant impact in computer vision and natural language processing. These datasets combined with fast computational resources have helped learn better models for various tasks in these domains.

2.2 Architectures

This section describes some commonly used deep learning architectures. These architectures are chosen based on the data or learning objective.

2.2.1 Deep Boltzmann Machines. Restricted Boltzmann Machines (RBM) [54] are undirected graphical models. They can be represented using bipartite graphs consisting of a single layer of observable variables $\mathbf{v}$ and a single layer of latent/hidden variables $\mathbf{h}$. RBMs are energy-based models with energy function

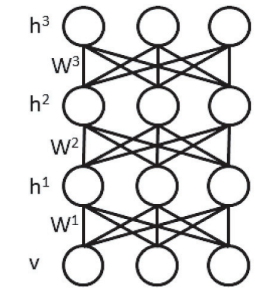
$$E(\boldsymbol{v}, \boldsymbol{h}) = -\mathbf{b}^T \boldsymbol{v} - \mathbf{c}^T \boldsymbol{h} - \boldsymbol{v}^T \mathbf{W} \boldsymbol{h}, \tag{1}$$

where $\mathbf{W}$ denotes edge weights between visible and hidden layers, and $\mathbf{b}$ and $\mathbf{c}$ are visible and hidden layer biases, respectively. The joint probability distribution is given by

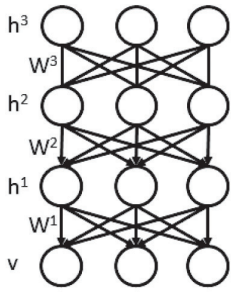
$$P(\mathbf{v} = \boldsymbol{v}, \mathbf{h} = \boldsymbol{h}) = \frac{1}{Z} exp\left(-E(\boldsymbol{v}, \boldsymbol{h})\right), \tag{2}$$

where Z is the partition function. With RBMs, $P(\mathbf{h}|\mathbf{v})$ and $P(\mathbf{v}|\mathbf{h})$ can be factored, thereby making computation and sampling easy. However, Z is still intractable. Hence, RBMs use approximate methods like Contrastive Divergence (CD) [15] for training. Deep Boltzmann Machines (DBM) [124] are constructed by stacking multiple RBMs. Figure 1(a) shows a DBM with three hidden layers.

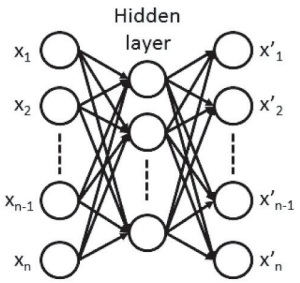
2.2.2 Deep Belief Networks. Deep Belief Networks (DBN) [55] are generative models and consist of one visible layer $\mathbf{v}$ and n hidden layers $\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(n)}$. A DBN is a mixed graphical model with directed edges between its lower layers (i.e., a Bayesian network) and undirected edges between its top two layers (i.e., an RBM). Figure 1(b) shows a DBN with three hidden layers. The



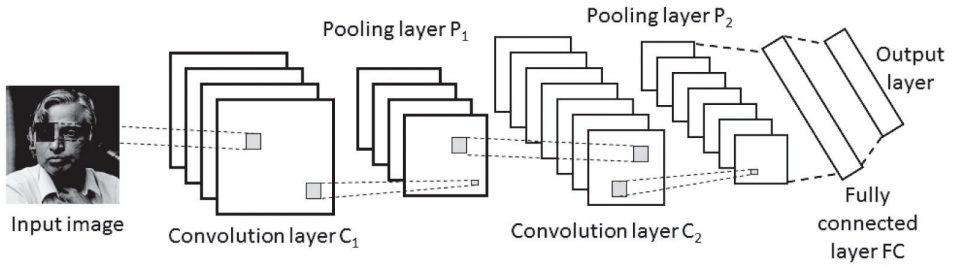
(a) Deep Boltzmann Machine



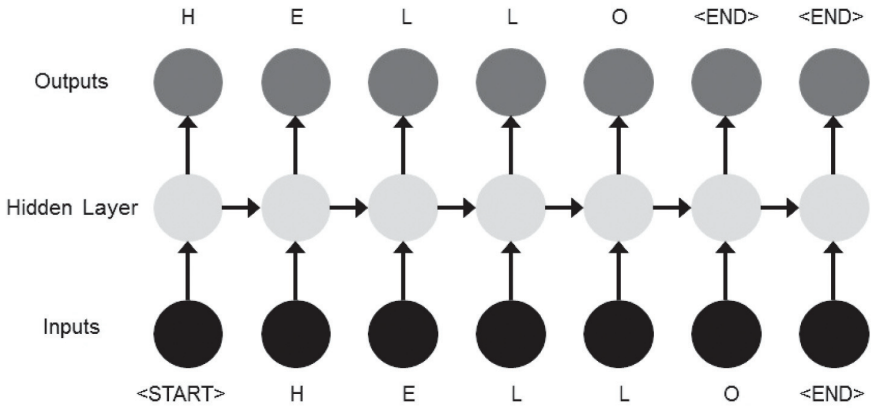
(b) Deep Belief Network



(c) Autoencoder



(d) Convolutional Neural Network (CNN)



(e) Recurrent Neural Network (RNN)

Fig. 1. Deep learning architectures.

joint probability distribution of DBN is given by

$$P\left(\mathbf{v}, \mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(n)}\right) = P\left(\mathbf{h}^{(n)}, \mathbf{h}^{(n-1)}\right) P\left(\mathbf{h}^{(n-2)} | \mathbf{h}^{(n-1)}\right) \ldots P\left(\mathbf{h}^{(1)} | \mathbf{h}^{(2)}\right) P\left(\mathbf{v} | \mathbf{h}^{(1)}\right).$$

A DBN with one hidden layer is just an RBM. Hence, to train a DBN with n hidden layers, we start with an RBM and train it using contrastive divergence method. Keeping the parameters of the first RBM fixed, the second RBM is trained using samples from the hidden layer of the first RBM. This procedure is repeated layer by layer n times. A trained DBN may be used to initialize a Multilayer

Perceptron (MLP), typically referred to as *pretraining*. The pretrained MLP may be further trained with additional data (i.e., *discriminative fine-tuning*).

2.2.3 Stacked Autoencoders. Autoencoders are widely used in unsupervised learning. These networks are trained to reconstruct their inputs under some constraints. Autoencoders are composed of two components: an encoder and a decoder. Given an input $\mathbf{x}$, the encoder learns an intermediate representation $\mathbf{h} = f(\mathbf{x})$ from which the input is reconstructed as $g(\mathbf{h}) \approx \mathbf{x}$ using the decoder. The encoder and decoder are jointly trained using backpropagation to minimize the reconstruction loss

$$\mathcal{L}(\mathbf{x}) = \|g(f(\mathbf{x})) - \mathbf{x}\|_2^2.$$

An autoencoder with one hidden layer is shown in Figure 1(c). Deep autoencoders can be constructed by adding more hidden layers to the encoder and decoder.

In order to prevent the autoencoder from learning an identity function, certain regularizations are imposed on $\mathbf{h}$. These regularizations encourage the autoencoder to learn useful properties in the data distribution. Depending on the regularization and cost function, autoencoders can be classified as sparse autoencoders [110], denoising autoencoders [10], or contractive autoencoders [120].

2.2.4 Convolutional Neural Networks. Convolutional Neural Networks (CNN) [77] are inspired by the mammalian visual cortex and are widely used in computer vision. As the name implies, *convolution* forms the primary operation of these networks. For example, given a 1D input $\mathbf{x}$ and a 1D kernel $\mathbf{k}$, the convolution operation is defined as

$$y[n] = (x * k)[n] = \sum_{m=-\infty}^{\infty} x[m]k[n-m].$$

In CNNs, convolutional layers perform convolution of the input with its kernels or weights. The convolved output is processed with a nonlinear activation function to produce a *feature map*. Unlike signal processing, these kernels are not predefined but learned during training. *Convolutional layers* require fewer parameters than their fully connected counterparts since they are characterized by sparse connections and weight sharing. Nodes in CNN layers are locally connected; that is, each unit in layer *k* receives input from a small neighborhood in layer *k-1*, also known as the *receptive field*. Furthermore, the kernel weights are shared across the entire image. By stacking multiple convolution layers, the higher level layers learn features from increasingly "global" receptive fields.

The convolution layers are interspersed with subsampling layers called *pooling layers*. Pooling layers take a small neighborhood of the feature map and replace it with some statistical information of the neighborhood (e.g., max pooling replaces a $n \times n$ neighborhood with the maximum activation in the neighborhood). Thus, pooling helps ensure invariance to shift, scaling, and distortion while reducing feature map dimensions. Deep CNNs are formed by alternately stacking convolution and pooling layers, as shown in Figure 1(d). One or more fully connected layers are appended to the last pooling layer to yield the final feature representation of the input.

2.2.5 Recurrent Neural Networks. Recurrent Neural Networks (RNN) [139] are widely used for processing sequential data like text, speech, and videos where data at every instant are dependent on previously encountered data. At a single timestep, RNNs consist of an input layer, one or more hidden layers, and an output layer. This network is unrolled for multiple timesteps such that hidden layers nodes of the previous timestep are connected to those of the current timestep. Thus, at every time step, hidden layer nodes receive two inputs, input data at that timestep and the hidden layer

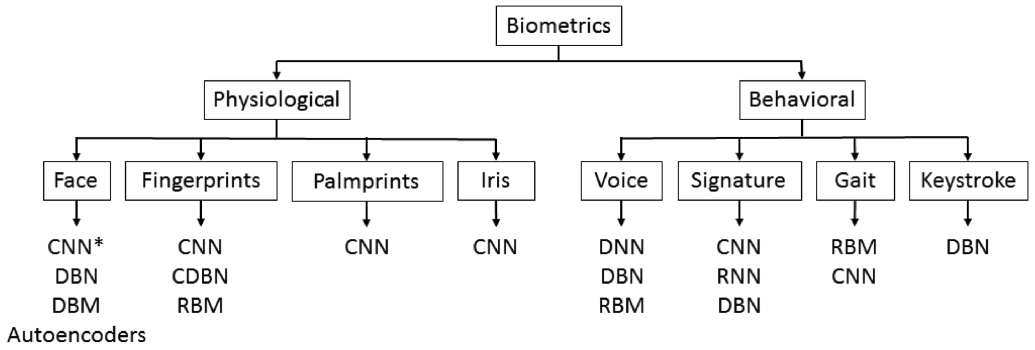


Fig. 2. Biometric modalities and corresponding deep learning architectures surveyed in this article.

representations from the previous timestep. Given an input $x^{(t)}$ at timestep t, the output $y^{(t)}$ is given by

$$
\begin{aligned}
m^{(t)} &= Ux^{(t)} + Wh^{(t-1)} + b \\
h^{(t)} &= tanh\left(m^{(t)}\right) \\
o^{(t)} &= Vh^{(t)} + c \\
y^{(t)} &= softmax\left(o^{(t)}\right),
\end{aligned}
$$

where U, W, and V are weight matrices connecting the input-to-hidden, hidden-to-hidden, and hidden-to-output layers and b, c are the corresponding bias vectors. An RNN with single hidden layer unrolled for seven timesteps is shown in Figure 1(e).

RNNs are considered to have a "memory" since the hidden layers encode information of the sequence encountered so far. This helps RNNs to encode long-range dependencies and hence provide better contextual reasoning. Practically, RNNs cannot be unrolled for too many steps due to vanishing gradients problem. Hence, architectures like Long Short-Term Memory (LSTMs) [56] were proposed to handle such challenges.

3 DEEP LEARNING IN BIOMETRICS

In this section, we survey deep learning approaches for the various biometric modalities shown in Figure 2. This includes four physiological biometrics (face, fingerprint, palmprint, and iris) and four behavioral biometrics (voice, signature, gait, and keystroke).

3.1 Face

Faces are the most common biometrics used by humans to recognize one another. Face recognition has varied uses including crime prevention, surveillance, forensic applications, and, more recently, in social networks.

Automatic face recognition, while nonintrusive, has various challenges associated with it due to imaging and physical factors. Some of these factors include illumination, pose, expression, occlusion, aging, facial style, and other physical factors. Though various deep learning approaches have been proposed specifically for face detection and alignment, we focus on face recognition and facial attribute estimation in this section.

3.1.1 Face Recognition. Face recognition can be formulated as a verification or identification problem. In verification mode, we verify whether a person is who he claims to be by comparing a

person's face to his previously collected gallery images. In identification mode, a person's face is compared with the gallery images of all individuals to establish a person's identity.

Automatic face recognition methods are broadly classified into feature-based approaches, which use local features, and appearance-based approaches, which use global representations. However, face recognition using a deep-learning framework hierarchically combines both local and global features while handling nuisance factors. Among different architectures that have been attempted for face recognition, CNN-based approaches have made a significant impact. Hence, in this section, we focus on CNN-based face recognition.

Following its success with object recognition, CNNs have been widely used for face recognition. However, CNNs trained for face recognition differ in their use of locally shared or unshared filters in the higher layers. This is due to the fact that different facial parts can be represented by different feature sets, and stationarity does not hold. An overview of CNN architectures used for face recognition is shown in Table 1.

With deep learning approaches, face identification and verification primarily differ in the cost function to be minimized. With face verification, one needs to deal with the similarity between two faces; hence this approach uses any variant of metric learning (e.g., Joint Bayesian, triplet loss) in the cost function to be minimized. However, face identification is a multiclass classification problem where one needs to predict the correct label out of many classes and includes a cross-entropy loss (with softmax layer) in its cost function.

A series of CNN architectures with varying objectives has been proposed by Sun et al. [135–137] for face verification. These methods involved learning *DeepID* features by training CNNs with joint identification-verification tasks. The learned features are used in conjunction with metric learning approaches like Joint Bayesian learning [16]. The approaches vary by the layers which are supervised by the joint identification-verification signal and by the CNN architectures. The authors made the following observations with these approaches: (i) training the network with more identities helps it learn identity-related features, which in turn helps improve verification accuracy; (ii) using overcomplete representation of faces at various scales and color channels seems to be effective; (iii) using an identification or a verification signal by itself is not optimal; and (iv) *DeepID* features are moderately sparse, selective to identity-related attributes (e.g., race, age, hair), and robust to occlusion. A few other approaches [18, 19, 84, 107] have also used similar techniques for face recognition (i.e., joint identification-verification cost function, very deep architectures with small filters, large training data, and metric learning methods).

Face verification performance has been pushed further by companies like Google and Facebook who are privy to large amounts of user photos. DeepFace [142], from Facebook, garnered much attention by proposing a face verification algorithm that was at par with human-level performance. It addressed alignment and face representation in the face recognition pipeline. To learn efficient face representations, a nine-layer CNN was trained on a large private database of 4 million Facebook images from 4,000 subjects. The authors further extended the DeepFace network [143] and inferred the following properties that affect the transferability of deep CNNs: (i) The bottleneck layer needs to be compact and acts as a regularizer for transfer learning; (ii) although performance saturates with increased training samples, it can be improved by using training samples selected by bootstrapping instead of random subsampling; and (iii) there is correlation between representation norm (measure of activated units in the final representation) and discriminability (i.e., lower representation norms have lower prediction confidences and higher prediction entropy). Similarly, Schroff et al. [128], from Google, proposed learning a direct embedding of images into a feature space using deep CNNs and triplet loss function (i.e., an image of a person must be closer to all other images of the same person than to any image of another person). The features thus learned

Table 1. Various CNN Architectures Used for Face Recognition

Method	Recognition mode	Input	#Convolution layers	#Pooling layers	#Fully connected layers	Classification	#nets
Sun et al. [136]	Verification	60 face patches (39×31 and 31×31 patches)	4 layers ($20 \times 4 \times 4$ SW, $40 \times 3 \times 3$ SW, $60 \times 3 \times 3$ LS, $80 \times 2 \times 2$ US)	3 (max pooling with 2×2 filters)	1 (160-dim DeepID features)	Joint Bayesian	60
Sun et al. [135]	Verification	25 face patches	4 layers ($20 \times 4 \times 4$ SW, $40 \times 3 \times 3$ SW, $60 \times 3 \times 3$ LS, $80 \times 2 \times 2$ LS)	3 (max pooling with 2×2 filters)	1 (160-dim DeepID2 features)	Joint Bayesian	25
Sun et al. [137]	Verification	25 face patches	4 layers ($128 \times 4 \times 4$ SW, $128 \times 3 \times 3$ SW, $128 \times 3 \times 3$ LS, $128 \times 2 \times 2$ LS)	3 (max pooling with 2×2 filters)	4 (512-dim DeepID2+ features)	Joint Bayesian	25
Hu et al. [59]	Verification - Medium	58×58 faces	3 layers ($16 \times 5 \times 5$, $32 \times 4 \times 4$, $48 \times 3 \times 3$)	3 layers (2×2 filters)	1 (160-dim)	Joint Bayesian	1
Taigman et al. [142]	Verification	152×152 RGB faces	5 layers ($32 \times 11 \times 11$ SW, $16 \times 9 \times 9$ SW, $16 \times 9 \times 9$ LS, $16 \times 7 \times 7$ LS, $16 \times 5 \times 5$ LS)	1 layer (3×3 filters)	1 (4096-dim)	Weighted χ^2 similarity	1
Zhu et al. [176]	Identification	96×96 grayscale faces	3 layers ($32 \times 5 \times 5$ LS, $32 \times 5 \times 5$ LS, $32 \times 5 \times 5$ LS)	2 layers (2×2 filters)	1 (96×96 reconstruction layer)	LDA	1
Pattabhi et al. [108]	Identification	28×32 faces	2 layers ($6 \times 5 \times 5$, $12 \times 5 \times 5$)	2 layers (2×2 filters)		Neural network	1
Liu et al. [84]	Verification & Identification	7 face patches	9 layers	Yes	1 layer	softmax layer & triplet loss	7
Zhou et al. [174]	Verification	4 face patches	10 layers	Yes	1 layer	softmax layer	4
Chen et al. [18]	Verification & Identification	100×100 faces	10 layers ($32 \times 3 \times 3$, $64 \times 3 \times 3$ (2), $128 \times 3 \times 3$, $96 \times 3 \times 3$, $192 \times 3 \times 3$, $128 \times 3 \times 3$, $256 \times 3 \times 3$, $160 \times 3 \times 3$, $320 \times 3 \times 3$)	5 layers (2×2 (4), 7×7 mean)	1 (10548-dim)	Joint Bayesian	1
Parkhi et al. [107]	Verification	224×224 face patches	13 layers ($64 \times 3 \times 3$ (2), $128 \times 3 \times 3$ (2), $256 \times 3 \times 3$ (3), $512 \times 3 \times 3$ (3))	5 layers (2×2 max-pooling)	3 (4096, 4096, 2622)	triplet loss metric learning	1
Schroff et al. [128]	Verification, identification & clustering	224×224 face patches	11 layers ($64 \times 7 \times 7$, $64 \times 1 \times 1$, $64 \times 3 \times 3$, $192 \times 1 \times 1$, $192 \times 3 \times 3$, $384 \times 1 \times 1$, $384 \times 3 \times 3$, $256 \times 1 \times 1$, $256 \times 3 \times 3$, $256 \times 1 \times 1$, $256 \times 3 \times 3$)	4 layers (3×3 filters)	3 (32×128, 32×128, 128)	triplet loss embedding	1
Wen et al. [153]	Verification	-	6 layers ($128 \times 3 \times 3$, $128 \times 3 \times 3$ (2), $128 \times 3 \times 3$, $256 \times 3 \times 3LS$, $256 \times 3 \times 3LS$, $256 \times 3 \times 3LS$)	4 layers (2×2)	1 (512-dim)	softmax & center loss	1
Sun et al. [138]	Verification	25 face patches	10 layers ($64 \times 3 \times 3$, $64 \times 3 \times 3$, $96 \times 3 \times 3$, $96 \times 3 \times 3$, $192 \times 3 \times 3$, $192 \times 3 \times 3$, $256 \times 3 \times 3$, $256 \times 3 \times 3$, $256 \times 3 \times 3LS$, $256 \times 3 \times 3LS$)	4 (max pooling with 2×2 filters)	1 (512-dim)	Joint Bayesian	25

$m \times n \times p$ in Convolutional Layers Indicate m Feature Maps, $n \times p$ filter size SW - shared weights, LS - locally shared weights, US - unshared weights.

can be used for face verification, identification, and clustering. They trained very deep CNN architectures [141, 165] with a private dataset of 100M–200M images of 8M subjects.

With the influx of various CNN architectures trained on large private datasets, it is difficult to understand the contribution of architecture design to performance improvement. To study the effectiveness of CNNs in face recognition using a common ground, Hu et al. [59] performed an evaluation of three CNN architectures (small, medium, and large), with all three networks trained on Labeled Faces in the Wild (LFW) dataset. They observed the following properties:

- Although color images contain more information, they do not contribute significantly to performance improvement.
- Normalizing learned features before classification improves performance.
- Dimensionality reduction of features (from 160-dim to 16-dim) still retains sufficient discriminative information for good performance.
- Features from higher level convolutional layers and softmax layers can be complementary to the typically used features from bottleneck layers.
- Multiple networks trained on various face parts help learn a powerful face representation as long as the face patches contain sufficient discriminative information.
- Metric learning methods, like Joint Bayesian, play an important role in improving face verification performance.

Zhu et al. [174] explored the effect of web-scale training data with a naïve 10-layer CNN trained using 5 million images of 20,000 identities. Following a critical analysis, they explain why performance improvement on standard benchmark datasets is far from real-world scenarios:

- While it is true that performance improves linearly with the number of identities used for training, using more identities with a lesser number of samples is not helpful. Most datasets gleaned from the Internet exhibit this long-tail property, with some identities having more samples while most others have few samples per identity.
- When training data increase, typically used tricks like Joint Bayesian, multistage features, clustering, and joint identification-verification do not contribute significantly.
- Most datasets contain images of celebrities—smiling, young, and beautiful with make-up. However, this does not reflect real-world scenarios, especially variations due to aging.

Face identification is a much more challenging issue since we have to handle a large number of classes. Zhu et al. [176] proposed Face Identity-Preserving (FIP) features to handle pose and illumination variations. This framework combines feature extraction layers with a reconstruction layer to reconstruct face images in a canonical view (i.e., frontal pose with neutral illumination). Pattabhi et al. [108] also propose a face identification system that learns efficient face representations under nonuniform illumination. Chiachia et al. [21] proposed learning a person-specific face representation using a three-layer CNN. Filters in the last layer were trained for person-specific representations using one-versus-all Support Vector Machine (SVM) training. Wen et al. [153] proposed using a novel center loss function along with softmax loss in CNNs to improve the interclass distances and intraclass compactness of face representations. The center loss function penalizes the distance between face representations and their corresponding class centers thereby yielding compact face representations for each identity. Masi et al. [3, 97] train an ensemble of pose-aware CNNs for face recognition where each CNN is trained for a specific pose using pose-specific images generated by 3D rendering.

Few approaches have been attempted in the effort to combat large labeled training datasets and associated training time. Sun et al. [138] attempt to iteratively train a sparse network from a denser model using correlations between neural activations of consecutive layers. The authors observe

that retaining only 12% of parameters can match the performance of DeepID2+ features, whereas retaining 26–76% of parameters surpasses the DeepID2+ performance. Peng et al. [109] attempt to reduce number of training images required to train a recognition network by jointly training a recognition network with a face alignment network. Reale et al. [117] evaluated the robustness of CNNs to mislabeled samples and limited training data. They observed that mislabeling adversely affects CNN performance compared to reducing training data.

Deep learning has also helped generate realistic face images using recent advances in generative models, and this has various applications, including domain-specific data augmentation. Tran et al. [144] use a hybrid discriminative-generative model such that the learned identity-specific latent representations can be used for pose-invariant face recognition and also to generate synthetic faces. The learned identity representations are disentanged from other variations like pose and hence can generate pose-specific face images of a person using a pose code representation. Yin et al. [163] used Generative Adversarial Networks (GANs) to generate facial images based on high-level attributes and also modified images using these attributes while maintaining identity.

3.1.2 Facial Attributes. Identifying facial attributes helps face recognition by narrowing down candidate matches. While previous methods have predominantly used hand-crafted texture descriptors, deep learning approaches have used learned features for identifying facial attributes like age, gender, and ethnicity.

Age estimation. Hierarchical representations from a CNN or an ensemble of CNNs have been used for age estimation via classification [62, 111] or regression [151]. Liu et al. [83] used a deep residual network to map representations of similar ages near and dissimilar ages beyond a margin using metric learning. The network also jointly learned to mine hard samples that satisfy these constraints.

Apparent age estimation requires crowdsourcing of age labels, and hence the benchmark datasets are usually small. Therefore, CNN-based approaches for apparent age estimation use features from models pretrained on ImageNet and later fine-tuned. Ranjan et al. [116] used features from a deep CNN trained for face identification [18] to perform age regression. Since training data may be skewed against young ($\leq$20) and old (>50) people, additional models were trained for the young and old to classify age hierarchically. Rothe et al. [121] fine-tuned a VGG-16 network for age estimation. The network is trained to classify images into multiple age groups, and an expectation of the final softmax probabilities is computed to estimate age. Liu et al. [87] used a joint regression and classification ensemble of eight CNNs for age estimation. The networks consisted of GoogleNet pretrained with face identities and then fine-tuned first with real age labels and then with apparent age labels. Similarly, Malli et al. [93] used an ensemble of three CNNs where each was trained to classify 34 different age groupings. The age groupings were different for each CNN to handle the difficult problem of classifying consequent ages. Antipov et al. [7] used an ensemble of 11 pretrained VGGNet models where labels were encoded using their training distributions. They augmented this setup with an ensemble of three CNNs fine-tuned specifically with images of children aged 0–12 to address another prominent age estimation problem, that of identifying children.

Gender estimation. Mansanet et al. [95] used DNNs trained on local face patches with high information content for gender recognition. It was observed that encoding location information of patches along with input helped improve classification. Juefei et al. [69] proposed an occlusion and low-resolution robust gender classification system by progressively training CNNs to emphasize the periocular region. This was done by training CNNs with sequentially blurred images while highlighting the periocular region. They observed that these blurred versions of images provide robustness to occlusion, noise, and low resolution. Jiang et al. [67] used high-level features from a five-layer CNN and low-level features from an autoencoder to perform gender estimation. Zhang

et al. [169] used an ensemble of 150 CNNs trained on poselets to mitigate pose and viewpoint effects in attribute (gender, hair style, clothes, etc.) estimation. Attributes cover only a small portion of an image, and it might be beneficial to use only certain parts that are pose-normalized (i.e., poselets) for feature learning. Zhang et al. [168] used a multitask framework to classify gender and smile using task-aware face cropping.

Multiattribute estimation. Samangouei et al. [126] proposed a multitask, parts-based CNN for estimating attributes to enable continuous mobile device authentication. They trained deep and wide variations of two CNNs: BinaryCNNs trained on a single attribute and MultiCNNs trained on multiple attributes. When sufficient training data were present, deep variants of MultiCNN performed best. Levi et al. [79] proposed a simple CNN architecture for age and gender classification since the number of target classes was lower. It was inferred that gender misclassifications mostly occurred with babies and young children who do not have well-developed gender characteristics. Li et al. [81] used shape information with kernel adaptation to handle the nonrigid nature of faces. The intuition was that fixed kernel functions will generate different activations for the same identity due to pose and expression variations. Hence, it would be beneficial to learn kernels that adapt to specific variations of an individual using a latent variable like face shape. Yi et al. [162] proposed a multitask, multiscale CNN for age, gender, and ethnicity estimation using local face patches in various scales. Outputs of all local CNNs were fused for use with a task-specific regression layer for age estimation and a softmax layer for gender and ethnicity classification. Liu et al. [88] performed attribute prediction using a two-stage cascaded CNN: LNet and ANet. LNet was pretrained with generic objects [26] for face localization and fine-tuned with attributes. ANet was pretrained with faces for identification and fine-tuned with attributes for learning extended attributes. The insights inferred from this approach are that (i) although LNet was fine-tuned with only attributes, its feature maps have strong responses for face locations; and (ii) ANet implicitly learns identity-related attributes like age, gender, and ethnicity even though it was pretrained for face identification. Srinivas et al. [132] used two CNNs trained on whole face and face patches for determining the age, gender, and ethnicity of East Asian subjects. Ranjan et al. [115] trained an all-in-one CNN multitask framework for various face tasks like face detection, pose estimation, alignment, recognition, and age and gender estimation. Similarly, a multitask framework using CNNs has proposed by Zhu et al. [175] and Han et al. [52] for facial attribute estimation.

3.2 Periocular Region

The periocular region (i.e., the region around the eye) is salient for face and facial attribute recognition and has proved to be useful in instances where the lower half of a face is occluded. Previous approaches [91, 100, 106] have used various hand-crafted texture descriptors to represent the periocular region. Instead, Nei et al. [105] used a two-layer convolutional RBM trained in an unsupervised manner using periocular image patches; the learned features were transformed using supervised metric learning to classify genuine/impostor pairs. Raghavendra et al. [112] used representations from autoencoders trained with texture features from 36 Maximum Response filters. Zhao et al. [172] used a multitask approach with three CNNs: One for periocular recognition and two others for auxiliary tasks of predicting gender and left/right eye. Ahuja et al. [4] used features from two CNN networks trained on faces and a root SIFT to perform iris and periocular recognition in visible spectrum.

3.3 Fingerprint

Fingerprints are the oldest known biometric modality. They consist of interleaved ridges and valleys formed by a combination of genetic and environmental factors. Features used for fingerprint

matching include global features (loop, delta, and whorl) and local features (minutiae). Fingerprint recognition can be challenging primarily due to large intraclass variations caused by displacement, nonlinear distortion, variable pressure, skin condition, and more.

Deep learning approaches have predominantly been used to extract both global and local features that aid fingerprint matching. Wang et al. [150] used features from a stacked autoecoder to classify fingerprints into arch, left/right loop, and whorl. Jiang et al. [66] used CNNs to extract minutiae from raw fingerprint images. They used two cascaded networks: JudgeNet with four CNNs trained on different scales to identify candidate patches containing minutiae and LocateNet to compute minutiae location while rejecting spurious patches. Su et al. [134] used CNNs to extract pores from raw fingerprint patches to aid automatic fingerprint identification. These are combined with minutiae and ridge patterns extracted using conventional approaches and fused using a unique matching scheme. Cao et al. [13] used minutiae features from ridge flow patterns using multiscale CNNs, a dictionary of ridge flow features, and texture-based features for automatic latent fingerprint recognition.

Some approaches have also been used to enhance and preprocess fingerprint images. Noisy fingerprint images require enhancement to improve clarity of ridge structures. Instead of conventionally used contextual filters, these approaches use filters learned from data for fingerprint enhancement. Sahasrabhude and Namboodiri [122] used continuous RBMs to correct distorted orientation fields of noisy images, a technique that was in turn used for fingerprint enhancement. In Sahasrabudhe and Namboodiri [123], the authors used a two-layer Convolutional DBN (CDBN) trained with clean fingerprint images having a significant variety of ridge orientations and frequencies. Noisy fingerprint images were reconstructed using inferences from the hidden layer representations of the CDBN. Deep learning has also been used to segment latent fingerprints. Ezeobiejesi et al. [35] pretrained a three-layer RBM with fingerprint image patches and fine-tuned the network to classify patches into those containing fingerprints or not.

3.4 Palmprint

The skin type on the palm's surface is quite similar to that found on fingertips, albeit with a larger surface area. Some commonly used palmprint features include shape, principal lines, wrinkles, delta points, and minutiae features.

Deep learning approaches have been primarily used to learn multiscale features for palmprint recognition. Jalali et al. [64] used a four-layer CNN trained with whole palmprint images and no palmprint Region-of-Interest (ROI) selection. Zhao et al. [171] used a two-layer RBM trained in an unsupervised fashion using 32×32 palmprint ROIs. Minaee and Wang [102] used a two-layer deep scattering CNN for palmprint recognition. Scattering networks are similar to CNNs, except that they use predefined wavelet transform filters rather than learning filters from data. Compared to this approach, Jalali et al. [64] did not use any palmprint ROIs, which demonstrates the robustness of CNNs for reasonable amounts of shift and distortion. Dian et al. [30] used Alexnet CNN with enhanced palmprint ROI images and Hausdorff distance for matching. Svoboda et al. [140] trained CNNs with palmprint ROIs and a d-prime loss function. They observed that d-prime loss based on genuine/impostor distributions worked much better than contrastive loss. In addition to recognition, CNNs have also been used for auxiliary tasks, such as palmprint ROI extraction. Bao et al. [8] use CNNs to identify the left or right hand and to detect key points on the palm for palmprint ROI extraction.

3.5 Iris

The iris is usually considered the most distinctive biometric trait. However, iris recognition systems, which depend on random texture information in the irises, usually present high

failure-to-acquire rate. This involves challenging preprocessing steps (e.g., iris segmentation, off-axis gaze correction, and removal of eyelashes). Daugman [24] proposed the groundbreaking approach for iris recognition used in many real-world applications today. This approach uses Gabor filters to capture the texture information of iris.

Deep learning approaches attempt to perform iris recognition by replacing the Gabor filters with filters learned using deep neural networks. Minnae et al. [101] used features extracted from VGG-Net with SVM. Liu et al. [86] used CNNs to learn source-specific filters for iris images from heterogeneous sources (i.e., visible and near-infrared). Gangwar et al. [41] used two very deep CNN architectures for iris recognition, one with eight convolutional layers and another with five convolution and two inception layers. The authors have demonstrated that this approach provides robustness with respect to segmentation and rotation/alignment errors.

Few attempts have been made to perform iris recognition in unconstrained conditions (e.g., with mobile devices). Raja et al. [114] used sparse autoencoders with images acquired from smartphones. Filters of sparse autoencoders trained on natural image patches were convolved with the segmented iris images. Local histograms of filter responses were used as feature representations. Similarly, Zhang et al. [170] used a fusion of hand-engineered ordinal measures features and learned features from a three-layer CNN for iris recognition on mobile devices owing to their complementary nature.

Deep learning approaches have also been used in related auxiliary tasks. Silva et al. [130] proposed iris image classification based on the characteristics of contact lenses. Du et al. [32] used a three-layer CNN to identify left and right iris images that have been incorrectly labeled.

3.6 Voice

With the increased use of mobile phones and VOIP technology, voice is perhaps the most accessible biometric modality. Voice generation consists of a physical part due to the low-level articulatory actions of an individual and a behavioral part due to environmental and sociolinguistic factors. Speaker recognition systems typically use spectral characteristics like Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction Cepstral Coefficients (PLCC) features. These systems can be categorized into text-dependent (i.e., depends on lexical content of an utterance) and text-independent systems (i.e., independent of lexical content), with latter being more challenging. Gaussian Mixture Models (GMMs) have been used widely to model the underlying distribution of speaker data, specifically to learn the Universal Background Model (UBM).

Deep learning approaches for speaker recognition are predominantly used to gather sufficient statistics instead of a GMM-UBM framework. Lei et al. [78] used DNNs trained for Automatic Speech Recognition (ASR) to compute frame alignments in contrast to the component Gaussians used in GMM-UBM. These frame posteriors and MFCC features were used to compute sufficient statistics for use with an i-vector/Probabilistic Linear Discriminant Analysis (PLDA) framework. This approach decouples features used for frame alignment and sufficient statistics allowing comparison of frames with the same phonetic content. Kenny et al. [70] used a similar approach, using DNNs to perform text-independent speaker recognition using Baum-Welch statistics. This frame representation captures both acoustic and phonetic events instead of only acoustic events, as in GMM-UBM systems. Compared with this approach, the method of Lei et al. [78] gave better performance, possibly due to more training data and richer input feature representation. Campbell et al. [12] used DBNs so that the output posterior probability can replace the GMM-UBM posterior probability commonly used in i-vector systems. Both pretrained and supervised DBNs were used while incorporating sparsity to ensure that the frame posteriors are localized. Garcia et al. [42] used DNNs trained on out-of-domain data to collect sufficient statistics. For the unsupervised domain adaptation task, these DNNs were used in conjunction with a PLDA framework where

out-of-domain PLDA parameters were adapted to in-domain data using hierarchical clustering. McLaren et al. [98] used CNNs to compute class posteriors for i-vector computation. CNNs were fed with the MFCC features of the target frame and context frames. Feature maps of different CNN filters were concatenated and fed to a seven-layer DNN to compute class posteriors.

Ghahabi et al. [44] proposed using generative Universal DBN (UDBN) and discriminatively trained DBNs for speaker verification. A Universal DBN was trained in an unsupervised fashion using i-vectors from different speakers. UDBN parameters were adapted to target and impostor samples using pretraining. Pretrained DBNs were fine-tuned discriminatively using label information for speaker verification. The authors further extended this system to multisession speaker recognition [45]. The authors also leverage the use of unlabeled data for i-vector–based speaker recognition [47]. Unlabeled data is used to build the UDBN, and DNNs are used to discriminatively model speakers by adapting from UDBN.

Some approaches use deep learning bottleneck layer outputs as speaker representations called d-vectors. Variani et al. [147] used a similar approach for text-dependent verification using smaller footprints. The averaged d-vectors of all frames in a speaker's utterances were used as a speaker-specific model. During the test phase, the d-vectors of an utterance were compared with the speaker-specific models. It was observed that d-vectors were less degraded by noise, and a fusion of i-vectors and d-vectors gave the best performance.

RBMs have been used in speaker recognition either to avoid or enhance the use of supervector/i-vector approaches. Vasilakakis et al. [148] used output distributions of stacked RBMs to learn pseudo-i-vectors for speaker recognition, though the approach fell short of standard i-vector performance. Stafylakis et al. [133] used RBMs as an alternative generative model to the conventional i-vector/PLDA model. They used a Siamese-twins model of RBMs with tied weights albeit with lower performance than an i-vector framework. Ghahabi et al. [46] used RBMs to produce nonlinear transformation and dimensionality reduction of GMM supervectors. Saleem et al. [125] used RBMs instead of UBMs to cluster acoustic features in an unsupervised manner and DNNs for discriminative prediction. They use contextual information from multiple frames, which has shown to improve performance.

Some approaches use DNNs to learn features invariant of different factors. Fu et al. [40] used features from three types of DNN for text-dependent recognition (unsupervised RBM, phone-discriminant DNN, and speaker-discriminant DNN). The features obtained from intermediate layers of these networks were used in tandem with PLCC features and a GMM-UBM framework. Yamada et al. [158] used bottleneck features of DNNs for distant-talking speaker identification. Under such scenarios, reverberation affects performance, and the authors hypothesized that DNNs may be able to learn a new feature space that retains speaker-discriminative characteristics while removing variations due to reverberations. Chen et al. [20] learned speaker-specific characteristics from MFCC features by training two deep subnets, a procedure that aims to reconstruct the input features with contrastive losses. Contrastive losses help capture speaker-specific characteristics while reconstruction losses help in regularization.

End-to-end verification systems have been proposed for both text-dependent and text-independent data. Heigold et al. [53] propose an integrated deep learning system that provides a verification score given few reference utterances and a test utterance. The authors demonstrated that this approach scales to large datasets by demonstrating their results on a private "Ok Google" dataset consisting of 80,000 speakers. It was also observed that an RNN-based system performed better than a DNN-based system, although with higher computation cost. Snyder et al. [131] used a similar approach for text-independent verification but with temporal pooling layers to handle variable-length inputs.

3.7 Signature

Signature is perhaps the second most widely used behavioral trait after voice biometrics. It has varied applications in legal, medical, and banking sectors. Signature verification methods can be online or offline. Online verification methods utilize the dynamics of the signing process (i.e., positions, trajectories, pressure, etc.). Offline methods use static images of signatures from which certain dynamic properties can be inferred. Signature verification can be challenging due to large intraclass variations and has low permanence.

Various deep learning approaches have been proposed for both online and offline signature verification. Drott et al. [31] used a three-layer CNN with sampled coordinates over recorded signature, pressure, angles, and velocities at these points as input. Fayazz et al. [37, 38] used a sparse autoencoder to learn local and global features in an unsupervised manner from normalized input image patches. The input image was convolved with the learned filters, pooled, and used as feature representations with SVM. Lai et al. [75] used RNNs to utilize the sequential nature of online verification. They trained a two-layer RNN with a Length Normalized Path Signature (LNPS) descriptor as input and triplet loss. The LNPS descriptor is designed to be scale- and rotation-invariant and can capture contextual information along with RNNs.

For offline verification, Hafeman et al. [51] used a two-stage approach using CNNs. In the writer-independent phase, feature representations were learned by training a CNN discriminatively to identify authors. Subsequently, these CNN features were used to train writer-dependent classifiers (SVMs) to distinguish between genuine and skilled impostors. They further evaluated this approach using different variants of AlexNet and VGG networks with four feature representation variants [50]. It was observed that while the model could separate signatures which looked completely different, it was susceptible to errors with slowly traced skilled forgeries. Ribeiro et al. [118] also followed a similar two-stage identification-verification approach using a three-layer DBN. The DBN was fed with a 179-dimensional input consisting of a Modified Direction Feature (MDF), width, a sixfold surface, and a wavelet transforms. Dey et al. [29] used a Siamese-twin network architecture with two identical AlexNet networks and contrastive loss for writer-independent verification.

3.8 Gait

Gait is a behavioral biometric that uses the shape and motion cues of a walking person to identify them. Shape features capture information during gait phases, while motion features capture information during the transition between these phases. Some of the challenges in gait recognition include variations in clothing, footwear, carrying objects, and walking speed. Gait recognition approaches typically use features computed from silhouette images due to the invariance of clothing color and texture.

Deep learning approaches for gait recognition are mostly silhouette shape-based and use Gait Energy Images (GEI) [94] or Chrono-Gait Images (CGI) [149] to learn features. Hossain et al. [58] used a RBM-based autoencoder to learn feature representations in an unsupervised manner from full-profile silhouette images. Zhang et al. [166] used pairs of GEIs as input to a Siamese-twin network and contrastive loss to embed gait representations of the same person nearby, irrespective of the view. During test time, the gait representation is attributed to the person whose training sample is nearest. Since cross-view gait images are taken into account during training, it performs well under cross-view scenarios. Yan et al. [159] used a three-layer CNN trained for multitask learning, where features learned from the GEI were used by top-level task-specific MLPs for gait recognition. It was observed that these methods do not generalize well to change of scenes or view and could be improved with larger training data.

Some approaches use pretrained or fine-tuned CNNs with GEIs to obtain feature representations for gait recognition [5, 129, 160]. Other approaches attempt to utilize temporal information in gait image sequences. Wu et al. [156] used a two-stream network and a 3-D CNN to utilize temporal information in gait sequences. They observed that 3D CNNs with temporal information gave significant improvement in performance. Li et al. [80] used gait features from a pretrained VGG-19 corresponding to all frames in a gait cycle and max-pooled to obtain a feature representation that captures spatiotemporal aspects. Wolf et al. [155] used 3D convolutions on input videos, where one channel consists of grayscale frames and other two channels contain optical flow information in X and Y directions. This enables the network to learn clothing-invariant representations.

3.9 Keystroke

Keystroke is a behavioral biometric trait that characterizes a person's typing pattern. Though uniqueness cannot be guaranteed, it might be beneficial for continuous verification of a person's identity after logging in. The primary challenges in keystroke recognition are due to large intra-class variations caused by typing behavior attributed to changes in mood, user position, keyboard, and the like.

Deep learning-methods use keystroke timing information to learn features for keystroke recognition. Deng et al. [27] compared both Deep Belief Nets and Gaussian Mixture Models for keystroke recognition. They used a two-layer DBN which was pretrained in an unsupervised manner using a 31-dim feature vector and fine-tuned with training data. It was observed that the DBN outperformed a GMM-based approach. The authors further extended this approach for mobile keystroke authentication [28] by using 4–35-dimensional features from various mobile sensors as input.

4 COMPARISON OF PUBLISHED RESULTS

4.1 Datasets

In this section, datasets that are commonly used for various biometric modalities are described.

4.1.1 Face.

- *Labeled Faces in the Wild (LFW)* [61]: This dataset is widely used for face recognition. It consists of 13,323 photos of 5,749 celebrities taken under unconstrained settings. For face verification, these images are organized as 6,000 face pairs in 10 splits. The dataset provides three benchmarking protocols: (a) *Image-restricted protocol*, (b) *unrestricted protocol*, (c) *unsupervised* protocol. Liao et al. [82] proposed the *BLUFR* protocol to exploit all the images in LFW while focusing on low False Alarm Rates (FAR).
- *YouTube Faces (YTF)* [154]: This dataset consists of 3,425 YouTube videos of 1,595 celebrities and is widely used for face recognition. The videos are organized as 5,000 video pairs in 10 splits.
- *AR Face database* [96]: This dataset consists of 4,000 frontal face images of 126 people under different illuminations, occlusions, and facial expressions. This dataset is used for face recognition and facial attribute recognition.
- *MORPH* [119]: This dataset, widely used for facial attribute estimation, consists of two albums of face images with age, gender, ethnicity, and other attributes. Album 1 consists of 1,724 images of 515 people. Album 2 consists of 55,134 images of 13,000 individuals.
- *IJB-A* [72]: This dataset consists of 5,396 images and 20,412 video frames from 500 people. This dataset is primarily used for face recognition.
- *Adience* [33]: This dataset is used for age and gender estimation using unconstrained facial images. It consists of 26,580 face images of 2,284 people with age labels grouped into eight categories.

- *ChaLearn 2015* [34]: This dataset is used to evaluate apparent age estimation. It consists of 2,476 training images and 1,136 validation images spanning ages 0 to 100.

4.1.2 Fingerprint.

- *Fingerprint Verification Competition (FVC 2002)* [92]: This dataset is widely used for fingerprint evaluation. The FVC 2002 challenge consists of three fingerprint datasets (DB1, DB2, and DB3) collected using different sensors. Each dataset consists of two sets: (i) Set A with 100 subjects and 8 impressions per subject, (ii) Set B with 10 subjects and 8 impressions per subject.
- *NIST SD27* [43]: This dataset consists of 258 latent fingerprints and corresponding reference fingerprints.
- *WVU DB* dataset: This consists of 449 latent fingerprints and corresponding reference fingerprints.

4.1.3 Palmprint. The *PolyU Hyperspectral Palmprint database* [167] consists of palm images from 190 volunteers imaged at 69 spectral bands. In total, the database consists of 5,240 images of 380 palms.

4.1.4 Iris.

- *VSSIRIS* [114]: This database consists of iris images acquired using iPhone 5S and Lumia 1020 under unconstrained conditions in visible spectrum. The dataset consists of 560 iris images acquired from 28 subjects, mostly from European countries.
- *Mobile Iris Challenge Evaluation (MICHE I)* [25]: MICHE I consists of iris images acquired under unconstrained conditions using smartphones. It consists of more than 3,732 images acquired from 92 subjects using three different smartphones.
- *Q-FIRE* [68]: The Q-FIRE dataset consists of iris images acquired at 5, 7, and 11 feet using the same sensor. The dataset consists of 3,123 high-resolution images acquired at 5ft and 2,902 low-resolution images acquired at 11ft from 160 subjects each.
- *LG2200 and LG4000* [2]: The ND-CrossSensor-Iris-2013 dataset consists of iris images taken with two iris sensors: LG2200 and LG4000. The LG2200 dataset consists of 116,564 iris images, and LG4000 consists of 29,986 iris images of 676 subjects.

4.1.5 Voice. The *NIST Speaker Recognition Evaluation (SRE)* [1] series provides a platform for improving research efforts in text-independent speaker recognition. NIST has been hosting the SRE series since 1996 and outlines an evaluation plan every year with training and testing tasks. Here, we list SREs that have been frequently used by deep learning approaches:

- *SRE 2012*: This SRE consists of nine distinct tests. Each test uses one of three training conditions (core, telephone, microphone) and one of five test conditions (core, extended, summed, known, and unknown). The benchmark uses 1,918 target speakers from speech corpora used in previous SREs.
- *SRE 2010*: This SRE also consists of nine distinct tests, where each test uses one of four training conditions (10-sec, core, 8conv, 8summed) and one of three test conditions (10-sec, core, summed).
- *SRE 2006*: This SRE consists of 15 distinct tests. The tests use one of five training conditions (10-sec, 5-min, 3conv, 8conv, 3summed) and one of four test conditions (10-sec, 5-min, summed, auxmic).

4.1.6 Signature.

- *GPDS-960 corpus* [146]: The GPDS-960 corpus consists of 24 genuine signatures and 30 forgeries of 960 individuals. A few variants include GDPS-160 and GDPS-300, which use genuine signatures and forgeries from 160 and 300 subjects respectively.
- *Signature verification competition (SVS 2004)* [161]: SVC 2004 consists of two datasets for two different verification tasks: for pen-based input devices like PDAs and for digitizing tablets. Each dataset consists of 100 sets of signatures with each set containing 20 genuine signatures and 20 skilled forgeries.

4.1.7 Gait.

- *CASIA-B* [164]: This dataset is widely used for gait recognition. The CASIA-B gait dataset consists of 3,718 videos of 101 subjects walking in straight lines with six different camera orientations. The subjects were instructed to walk under three different scenarios: normal walking, wearing a coat, and carrying a bag.
- *OU-ISIR LP dataset* [63]: The OU-ISIR dataset is the largest gait dataset, with two gait sequences each for 4,007 subjects. It comprises of four different view angles and no walking variations.

4.1.8 Keystroke.

CMU Benchmark Dataset [71]. The CMU keystroke dataset consists of keystroke dynamics for 51 subjects over eight sessions. The keystroke information includes dwell time for each key and latencies between two successive keys while typing a password string.

4.2 Results

In biometrics, deep learning has been explored to a great extent for face and speaker recognition. Very few approaches have evaluated deep learning techniques for other biometric modalities, thereby making it difficult to understand their implications with respect to these other modalities.

4.2.1 Face Recognition. For face verification, performance is measured using verification accuracy. Open-set or closed-set identification accuracy is used for face identification. Rank-1 identification accuracy is used for closed-set identification, and the Rank-1 Detection and Identification Rate (DIR) is used for open-set identification at a certain False Alarm Rate (FAR), typically 1%.

The reported results of various deep learning and other state-of-the-art approaches for LFW are listed in Table 2. For face verification, it can be seen that deep learning approaches outperform previous methods [14, 17, 89]. This suggests that deep learning approaches are able to better learn feature representations while accounting for various correlated factors. Many deep learning approaches [128, 137, 138, 153, 174] have surpassed human performance and are already close to 100%. In addition to their complex, deep architectures, high-performance metrics can be attributed to these approaches using images on an order of 0.2–200 million belonging to 10,000–8M identities for training. Further, Sun et al. [137] and Schroff et al. [128] use a joint identification-verification protocol, and multitask learning with the challenging identification task seems to help improve face verification performance as well. For face identification, Liu et al. [84] from Baidu Research have outperformed other deep learning approaches and Commercial Off-the-Shelf (COTS) software in both open- and closed-set identification tasks. This can be attributed to the deep nine-layer CNN trained on face patches from more than a million images and the use of triplet loss for metric learning, similar to Schroff et al. [128].

Another commonly used face verification dataset is YTF. Using the YTF dataset, the approaches of Parkhi et al. [107] and Schroff et al. [128] have outperformed other methods due to reasons mentioned earlier. The reported results of various approaches for the YTF and IJB-A datasets are

Table 2. Face Recognition Results: LFW (13,323 Images, 5,749 Celebrities)

Methods	Arch.	Training			Protocol	Verification	Open set	Closed
		Dataset	#img	#subj		acc (%)	acc (%)	acc (%)
Cao et al. [14]	Joint Bayesian	-	-	-	unrestricted	96.33 ± 1.08	-	-
Chen et al. [17]	LBP	-	-	-	unrestricted	95.17 ± 1.13	-	-
Lu et al. [89]	Gaussian Face	-	-	-	unrestricted	98.52 ± 0.66	-	-
Best et al. [11]	COTS-s1+s4	-	-	-	unrestricted	-	66.5	35
Sun et al. [136]	CNN*	CelebFaces+	202,599	10,177	unrestricted	97.45 ± 0.26	-	-
Sun et al. [135]	CNN*	CelebFaces+	202,599	10,177	unrestricted	96.39 ± 0.13	-	-
Sun et al. [137]	CNN*	CelebFaces+, WDRef	290,000	12,000	Jain	99.47 ± 0.12	80.7	95.0
Taigman et al. [142]	CNN*	SFC	4,000,000	4,000	unrestricted	97.35 ± 0.25	-	-
Liu et al. [84]	CNN*	Private	1,200,000	18,000	unrestricted	99.41	**95.80**	**98.03**
Chen et al. [18]	CNN*	CASIA-WebFace	490,356	10,548	unrestricted	97.45 ± 0.70	-	-
Parkhi et al. [107]	CNN*	Private	2,600,000	2,622	unrestricted	98.95	-	-
Schroff et al. [128]	CNN*	Private	100M-200M	8M	unrestricted	**99.63 ± 0.09**	-	-
Zhou et al. [174]	CNN*	Megvii	5,000,000	20,000	-	99.5	-	-
Taigman et al. [143]	CNN*	SFC	4,500,000	55,000	unrestricted	97.17	46.3	72.3
Wen et al. [153]	CNN*	private dataset	700,000	-	-	99.28	-	-
Sun et al. [138]	CNN*	CelebFaces+, WDRef	290,000	12,000	-	99.30	-	-
Peng et al. [109]	CNN*	CASIA WebFace	494,414	10,575	-	96.60	-	-

*Deep learning approach.

listed in Table 3. With the IJB-A dataset, it can be observed that a hybrid discriminative-generative DR-GAN [144] outperforms other discriminative CNN approaches for the challenging face identification task. This suggests that hybrid discriminative-generative approaches may be able to learn identity representations that are invariant to other nuisance factors.

4.2.2 Facial Attributes. Age estimation is evaluated using Mean Absolute Error (MAE) given by

$$MAE = \frac{\sum_{k=1}^{N} |\hat{y}_k - y_k|}{N}, \tag{3}$$

where y_k is the ground truth age of k^{th} sample, $\hat{y}_k$ is the corresponding predicted age, and N is the number of test samples.

Tables 4 and 5 list the reported results of various approaches for facial attribute estimation using different datasets. MORPH-II is the most common dataset for age and gender estimation. For age estimation using MORPH-II, deep learning approaches perform better than Biologically-Inspired Features (BIF) with Kernel Canonical Correlation Analysis (KCCA) and Kernel Partial Least Squares (KPLS). Apparent age estimation approaches [83, 121] seem to have the best performance across all datasets. Both approaches use pretrained deep networks (VGG-16 or GoogleNet), fine-tuned with age datasets. This seems to help in dealing with smaller age datasets. In addition, they also use CNN ensembles and a joint regression-classification framework for robust age estimation.

Table 3. Face Recognition Results: Other Datasets

Testing					Training				
Dataset	#img	#subj	Methods	Arch.	Dataset	#img	#subj	Ver. acc (%)	Id. acc (%)
YTF	3425 vid	1596	Cui et al. [23]	STFRD+ PMML	-	-	-	79.5 ± 2.5	-
			Mendez et al. [99]	VSOF+ OSS	-	-	-	79.7 ± 1.8	-
			Sun et al. [137]	CNN*	CelebFaces+, WDRef	290,000	12,000	93.2 ± 0.2	-
			Taigman et al. [142]	CNN*	SFC	4,000,000	4,000	91.4 ± 1.10	-
			Parkhi et al. [107]	CNN*	Private	2,600,000	2,622	**97.3**	-
			Schroff et al. [128]	CNN*	Private	100M-200M	8,000,000	95.12 ± 0.39	-
			Wen et al. [153]	CNN*	Private	700,000	-	94.9	-
			Sun et al. [138]	CNN*	CelebFaces+, WDRef	290,000	12,000	93.5	-
IJB-A	25808	500	Masi et al. [97]	CNN*	CASIA WebFace	500,000	-	82.6	94.6
			Sun et al. [138]	CNN*	CelebFaces+, WDRef	290,000	12,000	**92.7**	-
			Tran et al. [144]	DR-GAN*	CASIA WebFace, AFLW	520,407	10,575	83.1 ± 1.7	**95.3 ± 1.1**
			Ranjan et al. [115]	CNN*	CASIA WebFace	490,356	-	92.2	94.7

*Deep learning approach.

Gender estimation is evaluated using Accuracy (i.e., TP/N, where TP is the number of correct predictions and N is the number of test samples). With MORPH-II, the deep learning approaches [52, 81, 162] are comparable to other BIF approaches. Similar observations can be made from ethnicity estimation on MORPH-II.

4.2.3 Fingerprint. Fingerprint recognition results of deep learning and other state-of-the-art methods are listed in Table 6. The metric used for evaluation is Equal Error Rate (EER) and Rank-1 identification rate. Lower EER indicates better performance. It can be observed from the FVC 2002 results that the performance of deep learning methods is comparable to that of other state-of-the-art methods, with Short Time Fourier Transform (STFT) representation performing marginally better. With NIST SD27 and WVU DB, deep learning-based latent fingerprint recognition is on par or marginally better compared to a commercial fingerprint identification system.

4.2.4 Palmprint. Palmprint recognition methods were evaluated with the PolyU hyperspectral palmprint database [167]. The palmprint recognition results using deep learning techniques and previous state-of-the-art methods are listed in Table 7. It can be observed that recognition rates of previous methods were already high, and deep learning methods performed marginally better. As with fingerprint recognition, since the dataset used was small, effectiveness of deep learning methods, if any, remains inconclusive.

4.2.5 Iris. Iris recognition results of deep learning and other state-of-the-art methods for various datasets are listed in Table 8. Since some results were reported for various configurations, we have listed the mean results across all configurations for each approach. It can be observed from the results that deep learning methods outperform even Daugman's approach [24] in the visible

Table 4. Age Estimation Results

Attribute	Testing		Methods	Arch.	Training		#classes	MAE	Acc (%)
	Dataset	#img			Dataset	#img			
Age		-	Guo et al. [49]	BIF + KCCA	-	-	-	3.98	-
		-	Guo et al. [48]	BIF + KPLS	-	-	-	4.04	-
	MORPH-II	5-fold CV	Huerta et al. [62]	CNN*	MORPH-II	55,134	Regr.	3.88	-
		42,635	Yi et al. [162]	CNN*	MORPH-II	10,634	Regr.	3.63	-
		5,670	Qiu et al. [111]	CNN*	MORPH-II	47,582	Regr.	3.41	-
		44,634	Li et al. [81]	CNN*	MORPH-II	10,500	Regr.	3.61	-
		1,095	Wang et al. [151]	CNN*	MORPH-II	4,380	Regr.	4.77	-
		5-fold CV	Liu et al. [83]	CNN*	MORPH-II	-	Regr.	2.89	-
			Rothe et al. [121]	CNN*	IMDB-Wiki	523,051	101	**2.68**	-
			Han et al. [52]	CNN*	IMDB-Wiki	523,051	Regr	3.0	85.3
	Adience	5-fold CV	Levi et al. [79]	CNN*	Adience	26,000	8	-	84.70 ± 2.2
			Liu et al. [83]	CNN*	Adience, MORPH-II, ChaLearn	-	-	-	**98.2 ± 0.7**
			Rothe et al. [121]	CNN*	IMDB-Wiki	523,051	20	-	96.6 ± 0.9
	ChaLearn	1136	Liu et al. [83]	CNN*	MORPH-II, FG-Net	-	-	0.315	-
			Ranjan et al. [116]	CNN*	ChaLearn, Adience, MORPH	7,000	Regr	0.359	-
			Rothe et al. [121]	CNN*	IMDB-Wiki	523,051	101	**0.282**	-
			Liu et al. [87]	CNN*	CASIA WebFace, MORPH-II	1,315,000	-	0.287	-
			Ranjan et al. [115]	CNN*	IMDB-Wiki, Adience, MORPH	299,818	Regr	0.293	-
			Han et al. [52]	CNN*	IMDB-Wiki	523,051	3	0.289	-

CV, cross-validation; LOPO, leave one person out.
*Deep learning approach.

spectrum. These results show promise since the datasets are reasonably large, with the exception of VSSIRIS. Further, these datasets (except LG2200 and LG4000) consist of in-the-wild iris images taken under unconstrained conditions with smartphones and low-resolution images captured at varying distances. Thus, deep learning methods seem to provide an advantage for iris recognition under nonideal imaging conditions and in visible spectrum.

4.2.6 Voice. Despite having standard evaluation benchmarks like NIST SRE, none of the speaker recognition approaches we surveyed followed the protocol mentioned in the SRE evaluation plan. The approaches were each evaluated on different subsets of SRE benchmarks, hence we do not have a common ground on which to compare the performance of these papers. The results in Table 9 are based on the reported baseline performances for each of these approaches. A commonly used performance measure for speaker recognition is the EER. It can be observed from the results that performance of deep learning approaches is comparable to that of other GMM-based approaches. Since a common baseline was not used across these approaches, it is unclear whether deep learning approaches provide an advantage for speaker recognition.

Table 5. Gender and Ethnicity Estimation Results

Attribute	Testing		Methods	Arch.	Training		#classes	MAE	Acc (%)
	Dataset	#img			Dataset	#img			
Gender	MORPH-II	-	Guo et al. [49]	BIF + KCCA	-	-	-	-	98.45
		-	Guo et al. [48]	BIF + KPLS	-	-	-	-	98.35
		42,635	Yi et al. [162]	CNN*	MORPH-II	10,634	2	-	97.90
		44,634	Li et al. [81]	CNN*	MORPH-II	10,500	2	-	**98.48**
			Han et al. [52]	CNN*	IMDB-Wiki	523,051	2	-	98.0
	AR	1,275	Jiang et al. [67]	CNN*	FERET, CAS-PEAL	10,800	2	-	70.50
		3,288	Juefei et al. [69]	CNN*	MugshotDB, Pinellas	89,003	2	-	**85.62**
	Adience	5-fold CV	Levi et al. [79]	CNN*	Adience	26,000	2	-	86.80 ± 1.4
Ethnicity	MORPH-II	-	Guo et al. [49]	BIF + KCCA	-	-	-	-	98.95
		-	Guo et al. [48]	BIF + KPLS	MORPH-II	10,634	2	-	**99.0**
		42,635	Yi et al. [162]	CNN*	MORPH-II	10,634	2	-	98.60
			Han et al. [52]	CNN*	IMDB-Wiki	523,051	3	-	98.6

CV, cross-validation; LOPO, leave one person out.
*Deep learning approach.

Table 6. Fingerprint Recognition Results

Dataset	Methods	Representation	EER(%)	Rank1 (%)
FVC 2002	Hong et al. [57]	Gabor	24.34	-
	Chikkerur et al. [22]	STFT	**21.99**	-
	Sahasrabudhe et al. [122]	cRBM*	22.65	-
	Sahasrabudhe et al. [123]	cDBN*	23.95	-
NIST SD27	COTS latent AFIS	COTS	-	**67.0**
	CAO et al. [13]	CNN*	-	65.0
WVU DB	COTS latent AFIS	COTS	-	71.0
	CAO et al. [13]	CNN*	-	**75.0**

*Deep learning approach.

Table 7. Palmprint Recognition Results

Datasets	Methods	Representation	Recognition Accuracy(%)	EER(%)
PolyU	Lu et al. [90]	Enhanced GRCM	98.0	-
	Xu et al. [157]	Quaternion PCA+Quaternion DWT	98.83	-
	Jia et al. [65]	KPCA on HOL	99.73	-
	Jalali et al. [64]	CNN*	99.98	-
	Minaee et al. [102]	Scattering networks*	**100.0**	-
	Dian et al. [30]	CNN*	-	0.0443

*Deep learning approach.

Table 8. Iris Recognition Results

Dataset	Methods	Representation	EER(%)
MICHE-I	Daugman et al. [24]	Gabor	8.35
	Raghavendra et al. [113]	LBP	5.22
	Raja et al. [114]	SAE*	**3.93**
VSSIRIS	Daugman et al. [24]	Gabor	3.57
	Raghavendra et al. [113]	LBP	9.45
	Raja et al. [114]	SAE*	**1.70**
Q-FIRE	Weinberger et al. [152]	LMNN	1.73
	Liu et al. [85]	MDML	1.67
	Liu et al. [86]	CNN*	**0.15**
LG2200	Daugman et al. [24]	Gabor	7.12
	Gangwar et al. [41]	CNN*	**2.40**
LG4000	Daugman et al. [24]	Gabor	5.30
	Gangwar et al. [41]	CNN*	**1.82**

*Deep learning approach.

Table 9. Speaker Recognition Results

Testing					Training			
Dataset	#trials	#subj	Methods	Arch.	Dataset	#trials	#subj	EER(%)
SRE 2012	C2,C5	-	Li et al. [78]	UBM-EM (4096)	SRE 2012	-	1,918	2.18
				DNN*				**1.66**
	C2	1,040	Kenny et al. [70]	DNN*	SRE 2012	4,432	1,040	2.16
SRE 2012 training data	male	1,000	Vasilakakis et al. [148]	GMM	SRE 2012 training data	28,920	1,818	**0.45**
				DBN*				0.58
SRE 2010 telephone	-	7,196	Garcia et al. [42]	GMM	Switchboard I & II	33,039	3,114	6.92
				DNN*				4.20
			Saleem et al. [125]	DNN*	SRE 2004 & 2005 & 2006			**2.18**
SRE 2006	51,068	816	Ghahabi et al. [44]	i-vector	SRE 2004 & 2005	6,000	-	7.18
				DBN*				6.44
			Ghahabi et al. [46]	RBM*	SRE 2004 & 2005	6,125	-	7.58
			Ghahabi et al. [47]	DBN & DNN*	SRE 2004 & 2005 & 2006	-	-	**4.76**

*Deep learning approach.

4.2.7 Signature. Signature verification results for both online and offline verification are listed in Table 10. For online verification with the SVS2004 dataset, it can be observed that both deep learning and other approaches seem to perform comparably. This could be attributed to smaller dataset size. However, for offline verification with the GPDS-300 dataset, the results show that CNN

Table 10. Signature Verification Results

Datasets	Methods	Representation	EER(%)	FAR(%)	FRR(%)	Acc(%)
SVS 2004	Fallah et al. [36]	Mellin transform, MFCC, etc.	3.0	-	-	-
	Ansari et al. [6]	Fuzzy modeling	2.46	-	-	-
	Fayyaz et al. [38]	SAE*	**2.15**	-	-	-
	Lai et al. [75]	RNN*	2.37	-	-	-
GPDS-300	Ferrer et al. [39]	Geometric features	-	13.12	15.41	86.65
	Vargas et al. [145]	High-pressure pts	-	14.66	10.01	**87.67**
	Ribeiro et al. [118]	DBN*	-	14.67	20.25	82.85
	Hafemann et al. [51]	CNN*	10.70	9.08	20.60	-
	Haemann et al. [50]	CNN*	**3.47**	**5.13**	**6.55**	-
	Dey et al. [29]	CNN*	-	23.17	23.17	76.83

*Deep learning approach.

Table 11. Gait Recognition Results

Datasets	Methods	Representation	Accuracy(%)
CASIA-B	**Different views**		
	Kusakunniran et al. [74]	CCA	68.5
	Yu et al. [164]	GEI+NN	23.76
	Wu et al. [156]	CNN*	84.67
	Yan et al. [159]	CNN*	30.55
	Alotaibi et al. [5]	CNN*	85.51
	Hossain et al. [58]	RBM*	92.50
	Wolf et al. [155]	3D-CNN*	**97.35**
	Different scenes		
	Hu et al. [60]	LF+iHMM	71.76
	Kusakunniran et al. [73]	STIP	79.66
	Yan et al. [159]	CNN*	**95.0**
	Alotaibi et al. [5]	CNN*	86.70
OU-ISIR	Muramatsu et al. [103]	TCM+	72.80
	Muramatsu et al. [104]	wQVTM	70.51
	Wu et al. [156]	CNN*	94.8
	Zhang et al. [166]	CNN*	80.50
	Shiraga et al. [129]	CNN*	90.45
	Li et al. [80]	CNN*	**95.04**

*Deep learning approach.

representations [50] perform significantly better than other approaches. The two-phase writer-independent feature extraction and writer-dependent classification approach [50] seems to perform much better than the writer-independent Siamese-twin network approach [29], though both approaches used AlexNet models.

4.2.8 Gait. Gait recognition results for various methods are listed in Table 11 for the CASIA-B and OU-ISIR datasets. For CASIA-B, the results are reported for two conditions: probe and gallery under different views and scene conditions (with accessories like bag, coat etc.). For OU-ISIR, the results are reported for cross-view conditions. The results listed here are the average accuracies

Table 12. Keystroke Recognition Results on the CMU Keystroke Dataset

Methods	Representation	EER(%)
Zhong et al. [173]	I, II order statistics	8.4
Deng et al. [27]	GMM-UBM	5.5
Deng et al. [27]	DBN*	**3.5**

*Deep learning approach.

under various configurations. It can be observed from the results that deep learning methods seem to outperform other methods for different views and different scene conditions in both datasets. Using temporal information, either optical flow [155] or features of a gait cycle [80], seems to yield the best performance in both datasets.

4.2.9 Keystroke. Keystroke recognition results for various methods are listed in Table 12 for the CMU keystroke dataset [71]. The results show that the DBN-based approach [27] performs better than other methods. However, the dataset size is limited and does not necessarily imply the effectiveness of deep learning techniques.

5 DISCUSSION

In this section, we discuss how biometric recognition can be benefited by deep learning. Further, we also highlight potential gaps that exist between current approaches and their applicability to real-world applications.

5.1 Impact of Deep Learning in Biometrics

From Section 4.2, it is obvious that deep learning approaches have shown breakthrough performances primarily in face and speaker recognition. Extending this to other modalities, biometric recognition can be benefited by deep learning approaches due to the following factors:

(1) **Feature learning:** Deep learning methods have an edge over previous state-of-the-art methods owing to their ability to learn features from data. Biometric modalities like face and voice require both local and global features and are well-suited for hierarchical and compositional feature learning enabled by deep learning. Further, hand-crafting features for some modalities, like behavioral biometrics, gets abstract and learning features from raw data that will be useful for such modalities.
(2) **Invariant representations:** Since deep learning approaches learn features from the data, they disentangle correlated factors and learn feature representations that are robust to nuisance factors. This can be quite handy since real-word biometric data are quite noisy.
(3) **Generalization capability:** The learned features can be generalized to previously unseen datasets and also to other related tasks (e.g., features learned for face recognition can also be used for facial attribute estimation). In addition, pretraining improves feature learning by leveraging large amounts of unlabeled data when using smaller labeled training datasets.
(4) **Beyond bag-of-word features:** With increasing security and privacy concerns and an alarming increase of cybercrimes, researchers are exploring behavioral biometrics for authentication. With the resurgence of RNNs, the temporal aspects of such behavioral biometrics are captured efficiently compared to bag-of-word features.

5.2 Real-World Applicability

While deep learning has pushed the boundaries for some biometric modalities, there are a few gaps that need to be addressed:

(1) **Beyond face and voice recognition:** Deep learning in biometrics has been explored very little beyond face and speaker recognition. This could be attributed to widespread interest in face and speech information in other domains as well. Further, gathering face and speech data on a large scale is feasible with the advent of new technologies. However, the impact of deep learning on biometric recognition using other modalities is still unclear; for example, iris scans and fingerprints use only local edge-based features and may not benefit from deep learning.
(2) **Scaling up in terms of identification:** Authentication/verification been the primary focus of deep learning research in biometrics. Authentication is a relatively easy problem and scales well for a large number of subjects. However, the more challenging identification problem has received little attention so far. For large-scale identification, the biometric system needs to be able to distinguish between potentially millions of identities. This would require complex deep learning architectures that can capture subtle interclass differences and also handle large intraclass variability. Consequently, abundant amounts of training data would be required to capture these variations.
(3) **Large-scale datasets:** Though deep learning approaches have already surpassed human performance on some in-the-wild, large-scale datasets, these datasets do not meet the requirements of real-world, high-security applications. In addition, there is a lack of large-scale datasets for other modalities to benefit from deep learning. Even if large datasets are available, each individual needs to have a sufficient number of representative samples to account for various influencing factors.
(4) **Dataset quality:** Existing in-the-wild face recognition datasets are mostly celebrity images gleaned from the Internet. Though these images are unconstrained, they are still near-frontal, well-lit, high-resolution still images of people. Systems trained with such images may be unsuitable for real-world covert applications. For biometrics to benefit from deep learning, it is imperative to use large-scale datasets that capture real-world variations.
(5) **Computing resources:** With the increased use of mobile devices, secure authentication on such devices has become the need of the hour. However, if complex deep learning models are required for authentication, such devices might not have the necessary computing resources. Offloading authentication to a cloud-based system is also tricky since it involves transferring sensitive information and is vulnerable to attacks.
(6) **Training speed-up:** The success of deep learning has been largely demonstrated by industries with access to large amounts of data and computational resources. For most other researchers, computing resources are limited, and it is imperative to speed-up training of deep learning approaches. We need to strive for data-efficient learning algorithms.
(7) **Behavioral biometrics:** With increased online interaction in a connected world today, cyberspace authentication has begun to play an important role. Behavioral biometrics will be useful in such scenarios where physical biometrics wouldn't be as applicable. However, very few behavioral biometric modalities have been explored using deep learning approaches.

6 CONCLUSION

In this article, we have detailed the application of deep learning methods for various biometric modalities. Deep learning methods have shown remarkable improvement in the recognition of a

few biometric modalities (e.g., face and voice). These modalities have garnered much attention owing to their shared interests in other domains, the availability of large-scale datasets, and hierarchical feature representations. Previous state-of-the-art methods have been outperformed by deep learning methods due to their feature learning capability. Deep learning methods have also demonstrated good generalization capability for previously unseen datasets and related tasks.

In spite of the success with face and voice, deep learning approaches have been scarcely explored for other modalities. It is unclear whether these modalities can leverage the advantages of deep learning. This could be difficult either due to nature of the modality or availability of data.

Although deep learning research in biometrics is still nascent, we see potential in the following aspects:

- **Large-scale identification:** Though large-scale identification has not been explored extensively with deep learning approaches, they have the necessary framework to support it compared to the approaches used earlier. However, it would require complex models and abundant amounts of data and computational resources.
- **Behavioral biometrics:** With an increasing number of cybercrimes, use of behavioral biometrics is on the rise for authentication. RNNs have been shown to perform well for sequential data and would be beneficial for behavioral biometrics since they mostly consist of time-series data.
- **Robust to data noise:** Given representative real-word samples, deep learning methods can learn to disentangle various nuisance factors while learning discriminative feature representations.
- **Modeling biometric aging:** For biometric systems to be reliable for long periods of time, they must incorporate biometric aging. Generative deep learning approaches can be useful in this case by generating synthetically aged biometric data.
- **Biometric segmentation:** Since deep learning approaches learn from the data, they can also be used for some preprocessing tasks like segmenting the biometric data from a noisy background.
- **Fusion of multiple modalities:** Since deep learning approaches use neural network variants, it is possible to jointly train different architectures used by different biometric modalities. Hence, features of these modalities can be fused optimally for multibiometric applications.

REFERENCES

[1] 2012. NIST SRE Series. http://www.nist.gov/itl/iad/mig/sre.cfm.

[2] 2013. ND Cross-Sensor Iris Dataset. https://sites.google.com/a/nd.edu/public-cvrl/data-sets.

[3] Wael AbdAlmageed, Yue Wu, Stephen Rawls, Shai Harel, Tal Hassner, Iacopo Masi, Jongmoo Choi, Jatuporn Lekust, Jungyeon Kim, Prem Natarajan, and others. 2016. Face recognition using deep multi-pose representations. In *Proceedings of the IEEE 2016 Winter Conference on Applications of Computer Vision (WACV'16)*. IEEE, 1–9.

[4] Karan Ahuja, Rahul Islam, Ferdous A. Barbhuiya, and Kuntal Dey. 2016. A preliminary study of CNNs for iris and periocular verification in the visible spectrum. In *Proceedings of the 3rd International Conference on Pattern Recognition (ICPR'16)*. IEEE, 181–186.

[5] Munif Alotaibi and Ausif Mahmood. 2015. Improved gait recognition based on specialized deep convolutional neural networks. In *Proceedings of the 2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR'15)*. IEEE, 1–7.

[6] Abdul Quaiyum Ansari, Madasu Hanmandlu, Jaspreet Kour, and Abhineet Kumar Singh. 2013. Online signature verification using segment-level fuzzy modelling. *IET Biometrics* 3, 3 (2013), 113–127.

[7] Grigory Antipov, Moez Baccouche, Sid-Ahmed Berrani, and Jean-Luc Dugelay. 2016. Apparent age estimation from face images combining general and children-specialized deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 96–104.

[8] Xianjie Bao and Zhenhua Guo. 2016. Extracting region of interest for palmprint by convolutional neural networks. In *Proceedings of tth International Conference on Image Processing Theory Tools and Applications (IPTA'16)*. IEEE, 1–6.

[9] Yoshua Bengio. 2009. Learning deep architectures for AI. *Foundations and Trends® in Machine Learning* 2, 1 (2009), 1–127.
[10] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. 2013. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*. 899–907.
[11] Lacey Best-Rowden, Hu Han, Charles Otto, Brendan F. Klare, and Anil K. Jain. 2014. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2144–2157.
[12] William M. Campbell. 2014. Using deep belief networks for vector-based speaker recognition. In *Proceedings of the INTERSPEECH Conference*. 676–680.
[13] Kai Cao and Anil K. Jain. 2017. Automated latent fingerprint recognition. arXiv:1704.01925.
[14] Xudong Cao, David Wipf, Fang Wen, Genquan Duan, and Jian Sun. 2013. A practical transfer learning algorithm for face verification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3208–3215.
[15] Miguel A. Carreira-Perpinan and Geoffrey Hinton. 2005. On contrastive divergence learning. In *Proceedings of the 10th Workshop on Artificial Intelligence and Statistics (AISTATS'05)*. 33–40.
[16] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. 2012. Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*. Springer, 566–579.
[17] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. 2013. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3025–3032.
[18] Jun-Cheng Chen, Vishal M. Patel, and Rama Chellappa. 2016. Unconstrained face verification using deep CNN features. In *Proceedings of the IEEE 2016 Winter Conference on Applications of Computer Vision (WACV'16)*. IEEE, 1–9.
[19] Jun-Cheng Chen, Rajeev Ranjan, Amit Kumar, Ching-Hui Chen, Vishal M. Patel, and Rama Chellappa. 2015. 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*. 118–126.
[20] Ke Chen and Ahmad Salman. 2011. Learning speaker-specific characteristics with a deep neural architecture. *IEEE Transactions on Neural Networks* 22, 11 (2011), 1744–1756.
[21] Giovani Chiachia, Alexandre X. Falcao, Nicolas Pinto, Anderson Rocha, and David Cox. 2014. Learning person-specific representations from faces in the wild. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2089–2099.
[22] Sharat Chikkerur, Alexander N. Cartwright, and Venu Govindaraju. 2007. Fingerprint enhancement using STFT analysis. *Pattern Recognition* 40, 1 (2007), 198–211.
[23] Zhen Cui, Wen Li, Dong Xu, Shiguang Shan, and Xilin Chen. 2013. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3554–3561.
[24] John Daugman. 2004. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology* 14, 1 (2004), 21–30.
[25] Maria De Marsico, Michele Nappi, Daniel Riccio, and Harry Wechsler. 2015. Mobile iris challenge evaluation (MICHE)-I, biometric iris dataset and protocols. *Pattern Recognition Letters* 57 (2015), 17–23.
[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, 248–255.
[27] Yunbin Deng and Yu Zhong. 2013. Keystroke dynamics user authentication based on Gaussian mixture model and deep belief nets. *ISRN Signal Processing* 2013, Article 565183, 7 pages.
[28] Yunbin Deng and Yu Zhong. 2015. Keystroke dynamics advances for mobile devices using deep neural network. *Gate to Computer Science and Research* 2, 59.
[29] Sounak Dey, Anjan Dutta, J. Ignacio Toledo, Suman K. Ghosh, Josep Lladós, and Umapada Pal. 2017. SigNet: Convolutional Siamese network for writer independent offline signature verification. arXiv:1707.02131.
[30] Liu Dian and Sun Dongmei. 2016. Contactless palmprint recognition based on convolutional neural network. In *Proceedings of the IEEE 13th International Conference on Signal Processing (ICSP'16)*. IEEE, 1363–1367.
[31] Beatrice Drott and Thomas Hassan-Reza. 2015. On-line handwritten signature verification using machine learning techniques with a deep learning approach. Master's Theses in Mathematical Sciences.
[32] Yixin Du, Thirimachos Bourlai, and Jeremy Dawson. 2016. Automated classification of mislabeled near-infrared left and right iris images using convolutional neural networks. In *Proceedings of the IEEE 8th International Conference on Biometrics Theory, Applications, and Systems (BTAS'16)*. IEEE, 1–6.
[33] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2170–2179.
[34] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzalez, Hugo J. Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. 2015. ChaLearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1–9.

[35] Jude Ezeobiejesi and Bir Bhanu. 2017. Latent fingerprint image segmentation using deep neural network. In *Deep Learning for Biometrics*. Springer, 83–107.
[36] Asghar Fallah, Mahdi Jamaati, and Ali Soleamani. 2011. A new online signature verification system based on combining Mellin transform, MFCC and neural network. *Digital Signal Processing* 21, 2 (2011), 404–416.
[37] Mohsen Fayyaz, Mohammad Hajizadeh_Saffar, Mohammad Sabokrou, and Mahmood Fathy. 2015a. Feature representation for online signature verification. arXiv:1505.08153.
[38] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Hoseini, and M. Fathy. 2015b. Online signature verification based on feature representation. In *Proceedings of the International Symposium on Artificial Intelligence and Signal Processing (AISP'15)*. IEEE, 211–216.
[39] Miguel A. Ferrer, Jesus B. Alonso, and Carlos M. Travieso. 2005. Offline geometric parameters for automatic signature verification using fixed-point arithmetic. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 6 (2005), 993–997.
[40] Tianfan Fu, Yanmin Qian, Yuan Liu, and Kai Yu. 2014. Tandem deep features for text-dependent speaker verification. In *Proceedings of INTERSPEECH Conference*. 1327–1331.
[41] Abhishek Gangwar and Akanksha Joshi. 2016. DeepIrisNet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'16)*. IEEE, 2301–2305.
[42] Daniel Garcia-Romero, Xiaohui Zhang, Alan McCree, and Daniel Povey. 2014. Improving speaker recognition performance in the domain adaptation challenge using deep neural networks. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT'14)*. IEEE, 378–383.
[43] Michael D. Garris and R. Michael McCabe. 2000. *NIST Special Database 27: Fingerprint Minutiae from Latent and Matching Tenprint Images*. Technical Report 6534. NIST.
[44] Omid Ghahabi and Javier Hernando. 2014a. Deep belief networks for i-Vector based speaker recognition. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*. IEEE, 1700–1704.
[45] Omid Ghahabi and Javier Hernando. 2014b. i-Vector modeling with deep belief networks for multi-session speaker recognition. *Network* 20 (2014), 13.
[46] Omid Ghahabi and Javier Hernando. 2015. Restricted Boltzmann machine supervectors for speaker recognition. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'15)*. IEEE, 4804–4808.
[47] Omid Ghahabi and Javier Hernando. 2017. Deep learning backend for single and multisession i-Vector speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 4 (2017), 807–817.
[48] Guodong Guo and Guowang Mu. 2011. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, 657–664.
[49] Guodong Guo and Guowang Mu. 2013. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In *Proceedings of the IEEE 10th International Conference and Workshops on Automatic Face and Gesture Recognition (FG'13)*. IEEE, 1–6.
[50] Luiz G. Hafemann, Robert Sabourin, and Luiz S. Oliveira. 2016b. Analyzing features learned for offline signature verification using deep CNNs. In *Proceedings of the IEEE 23rd International Conference on Pattern Recognition (ICPR'16)*. IEEE, 2989–2994.
[51] Luiz G. Hafemann, Robert Sabourin, and Luiz S. Oliveira. 2016a. Writer-independent feature learning for offline signature verification using deep convolutional neural networks. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'16)*. IEEE, 2576–2583.
[52] Hu Han, Anil K. Jain, Shiguang Shan, and Xilin Chen. 2017. Heterogeneous face attribute estimation: A deep multitask learning approach. arXiv:1706.00906.
[53] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. 2016. End-to-end text-dependent speaker verification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16)*. IEEE, 5115–5119.
[54] Geoffrey Hinton. 2010. A practical guide to training restricted Boltzmann machines. *Momentum* 9, 1 (2010), 926.
[55] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18, 7 (2006), 1527–1554.
[56] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[57] Lin Hong, Yifei Wan, and Anil Jain. 1998. Fingerprint image enhancement: Algorithm and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8 (1998), 777–789.
[58] Emdad Hossain and Girija Chetty. 2013. Multimodal feature learning for gait biometric based human identity recognition. In *Proceedings of the International Conference on Neural Information Processing*. 721–728.

[59] Guosheng Hu, Yongxin Yang, Dong Yi, Josef Kittler, William Christmas, Stan Li, and Timothy Hospedales. 2015. 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*. 142–150.

[60] Maodi Hu, Yunhong Wang, Zhaoxiang Zhang, De Zhang, and James J. Little. 2013. Incremental learning for video-based gait recognition with LBP flow. *IEEE Transactions on Cybernetics* 43, 1 (2013), 77–89.

[61] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report 07-49, University of Massachusetts, Amherst.

[62] Ivan Huerta, Carles Fernández, Carlos Segura, Javier Hernando, and Andrea Prati. 2015. A deep analysis on age estimation. *Pattern Recognition Letters* 68 (2015), 239–249.

[63] Haruyuki Iwama, Mayu Okumura, Yasushi Makihara, and Yasushi Yagi. 2012. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security* 7, 5 (2012), 1511–1521.

[64] Amin Jalali, Rommohan Mallipeddi, and Minho Lee. 2015. Deformation invariant and contactless palmprint recognition using convolutional neural network. In *Proceedings of the 3rd International Conference on Human-Agent Interaction*. ACM, 209–212.

[65] Wei Jia, Rong-Xiang Hu, Ying-Ke Lei, Yang Zhao, and Jie Gui. 2014. Histogram of oriented lines for palmprint recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, 3 (2014), 385–395.

[66] Lu Jiang, Tong Zhao, Chaochao Bai, A. Yong, and Min Wu. 2016. A direct fingerprint minutiae extraction approach based on convolutional neural networks. In *Proceeedings of the 2016 International Joint Conference on Neural Networks (IJCNN'16)*. IEEE, 571–578.

[67] Yuxin Jiang, Songbin Li, Peng Liu, and Qiongxing Dai. 2014. Multi-feature deep learning for face gender recognition. In *Proceedings of the IEEE 7th Joint International Information Technology and Artificial Intelligence Conference (ITAIC'14)*. IEEE, 507–511.

[68] P. A. Johnson, P. Lopez-Meyer, N. Sazonova, F. Hua, and S. Schuckers. 2010. Quality in face and iris research ensemble (Q-FIRE). In *Proceedings of the IEEE 4th International Conference on Biometrics: Theory Applications and Systems (BTAS'10)*. IEEE, 1–6.

[69] Felix Juefei-Xu, Eshan Verma, Parag Goel, Anisha Cherodian, and Marios Savvides. 2016. DeepGender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 68–77.

[70] Patrick Kenny, Vishwa Gupta, Themos Stafylakis, P. Ouellet, and J. Alam. 2014. Deep neural networks for extracting Baum-Welch statistics for speaker recognition. In *Proceedings of the Odyssey Conference*. 293–298.

[71] Kevin S. Killourhy and Roy A. Maxion. 2009. Comparing anomaly-detection algorithms for keystroke dynamics. In *Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'09)*. IEEE, 125–134.

[72] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K. Jain. 2015. 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*. 1931–1939.

[73] Worapan Kusakunniran. 2014. Recognizing gaits on spatio-temporal feature domain. *IEEE Transactions on Information Forensics and Security* 9, 9 (2014), 1416–1423.

[74] Worapan Kusakunniran, Qiang Wu, Jian Zhang, Hongdong Li, and Liang Wang. 2014. Recognizing gaits across views through correlated motion co-clustering. *IEEE Transactions on Image Processing* 23, 2 (2014), 696–709.

[75] Songxuan Lai, Lianwen Jin, and Weixin Yang. 2017. Online signature verification using recurrent neural network and length-normalized path signature. arXiv:1705.06849.

[76] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436.

[77] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 11 (1998), 2278–2324.

[78] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*. IEEE, 1695–1699.

[79] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 34–42.

[80] Chao Li, Xin Min, Shouqian Sun, Wenqian Lin, and Zhichuan Tang. 2017. DeepGait: A learning deep convolutional representation for view-invariant gait recognition using joint Bayesian. *Applied Sciences* 7, 3 (2017), 210.

[81] Shaoxin Li, Junliang Xing, Zhiheng Niu, Shiguang Shan, and Shuicheng Yan. 2015. Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 222–230.

[82] Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z. Li. 2014. A benchmark study of large-scale unconstrained face recognition. In *IEEE International Joint Conference on Biometrics (IJCB 2014)*. IEEE, 1–8.

[83] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2017. Label-sensitive deep metric learning for facial age estimation. *IEEE Transactions on Information Forensics and Security* 13, 2, 292–305.
[84] Jinguo Liu, Yafeng Deng, and Chang Huang. 2015a. Targeting ultimate accuracy: Face recognition via deep embedding. arXiv:1506.07310.
[85] Jing Liu, Zhenan Sun, and Tieniu Tan. 2014. Distance metric learning for recognizing low-resolution iris images. *Neurocomputing* 144 (2014), 484–492.
[86] Nianfeng Liu, Man Zhang, Haiqing Li, Zhenan Sun, and Tieniu Tan. 2015. DeepIris: Learning pairwise filter bank for heterogeneous iris verification. *Pattern Recognition Letters* 82, 2, 154–161.
[87] Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. 2015b. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 16–24.
[88] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*. 3730–3738.
[89] Chaochao Lu and Xiaoou Tang. 2015. Surpassing human-level face verification performance on LFW with GaussianFace. In *Proceedings of the AAAI Conference*. 3811–3819.
[90] J. Lu, Y. Zhao, and J. Hu. 2009. Enhanced Gabor-based region covariance matrices for palmprint recognition. *Electronics Letters* 45, 17 (2009), 880–881.
[91] Jamie R. Lyle, Philip E. Miller, Shrinivas J. Pundlik, and Damon L. Woodard. 2010. Soft biometric classification using periocular region features. In *Proceedings of the IEEE 4th International Conference on Biometrics: Theory Applications and Systems (BTAS'10)*. IEEE, 1–7.
[92] Dario Maio, Davide Maltoni, Raffaele Cappelli, James L. Wayman, and Anil K. Jain. 2002. FVC2002: Second fingerprint verification competition. In *Proceedings of the 16th International Conference on Pattern Recognition*. Vol. 3. IEEE, 811–814.
[93] Refik Can Malli, Mehmet Aygün, and Hazim Kemal Ekenel. 2016. Apparent age estimation using ensemble of deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'16)*. IEEE, 714–721.
[94] Ju Man and Bir Bhanu. 2006. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 2 (2006), 316–322.
[95] Jordi Mansanet, Alberto Albiol, and Roberto Paredes. 2016. Local deep neural networks for gender recognition. *Pattern Recognition Letters* 70 (2016), 80–86.
[96] Aleix M. Martinez. 1998. *The AR Face Database*. Technical Report 24. CVC.
[97] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. 2016. Pose-aware face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4838–4846.
[98] Mitchell McLaren, Yun Lei, Nicolas Scheffer, and Luciana Ferrer. 2014. Application of convolutional neural networks to speaker recognition in noisy conditions. In *Proceedings of the INTERSPEECH Conference*. 686–690.
[99] Heydi Méndez-Vázquez, Yoanna Martínez-Díaz, and Zhenhua Chai. 2013. Volume structured ordinal features with background similarity measure for video face recognition. In *Proceedings of the International Conference on Biometrics (ICB'13)*. IEEE, 1–6.
[100] Philip E. Miller, Jamie R. Lyle, Shrinivas J. Pundlik, and Damon L . Woodard. 2010. Performance evaluation of local appearance based periocular recognition. In *Proceedings of the IEEE 4th International Conference on Biometrics: Theory Applications and Systems (BTAS'10)*. IEEE, 1–6.
[101] Shervin Minaee, Amirali Abdolrashidiy, and Yao Wang. 2016. An experimental study of deep convolutional features for iris recognition. In *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB'16)*. IEEE, 1–6.
[102] Shervin Minaee and Yao Wang. 2016. Palmprint recognition using deep scattering convolutional network. arXiv:1603.09027.
[103] Daigo Muramatsu, Yasushi Makihara, and Yasushi Yagi. 2015. Cross-view gait recognition by fusion of multiple transformation consistency measures. *IET Biometrics* 4, 2 (2015), 62–73.
[104] Daigo Muramatsu, Yasushi Makihara, and Yasushi Yagi. 2016. View transformation model incorporating quality measures for cross-view gait recognition. *IEEE Transactions on Cybernetics* 46, 7 (2016), 1602–1615.
[105] Lei Nie, Ajay Kumar, and Song Zhan. 2014. Periocular recognition using unsupervised convolutional RBM feature learning. In *Proceedings of the ICPR Conference*. 399–404.
[106] Unsang Park, Arun Ross, and Anil K. Jain. 2009. Periocular biometrics in the visible spectrum: A feasibility study. In *Proceedings of the IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS'09)*. IEEE, 1–6.
[107] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, Vol. 1. 6.

[108] N. Pattabhi Ramaiah, Earnest Paul Ijjina, and C. Krishna Mohan. 2015. Illumination invariant face recognition using convolutional neural networks. In *Proceedings of the IEEE International Conference on Signal Processing, Informatics, Communication, and Energy Systems (SPICES'15)*. IEEE, 1–4.

[109] Xi Peng, Nalini Ratha, and Sharathchandra Pankanti. 2016. Learning face recognition from limited training data using deep neural networks. In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR'16)*. IEEE, 1442–1447.

[110] Christopher Poultney, Sumit Chopra, and Yann L. Cun 2006. Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems*. 1137–1144.

[111] Jiayan Qiu, Yuchao Dai, Yuhang Zhang, and Jose M. Alvarez. 2015. Hierarchical aggregation based deep aging feature for age prediction. In *Proceedings of the 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA'15)*. IEEE, 1–5.

[112] Ramachandra Raghavendra and Christoph Busch. 2016. Learning deeply coupled autoencoders for smartphone based robust periocular verification. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP'16)*. IEEE, 325–329.

[113] Ramachandra Raghavendra, Kiran B. Raja, Bian Yang, and Christoph Busch. 2013. Combining iris and periocular recognition using light field camera. In *Proceedings of the 2013 2nd IAPR Asian Conference on Pattern Recognition (ACPR'13)*. IEEE, 155–159.

[114] Kiran B. Raja, Ramachandra Raghavendra, Vinay Krishna Vemuri, and Christoph Busch. 2015. Smartphone based visible iris recognition using deep sparse filtering. *Pattern Recognition Letters* 57 (2015), 33–42.

[115] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa. 2017. An all-in-one convolutional neural network for face analysis. In *Proceedings of the IEEE 12th International Conference on Automatic Face and Gesture Recognition (FG'17)*. IEEE, 17–24.

[116] Rajeev Ranjan, Sabrina Zhou, Jun Cheng Chen, Amit Kumar, Azadeh Alavi, Vishal M. Patel, and Rama Chellappa. 2015. Unconstrained age estimation with deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 109–117.

[117] Christopher Reale, Nasser M. Nasrabadi, and Rama Chellappa. 2016. An analysis of the robustness of deep face recognition networks to noisy training labels. In *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP'16)*. IEEE, 1192–1196.

[118] Bernardete Ribeiro, Ivo Gonçalves, Sérgio Santos, and Alexander Kovacec. 2011. Deep learning networks for off-line handwritten signature recognition. In *Proceedings of the Iberoamerican Congress on Pattern Recognition*. 523–532.

[119] Karl Ricanek and Tamirat Tesafaye. 2006. Morph: A longitudinal image database of normal adult age-progression. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'06)*. IEEE, 341–345.

[120] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. 2011. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 833–840.

[121] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision* 126, 2-4, 144–157.

[122] Mihir Sahasrabudhe and Anoop M. Namboodiri. 2013. Learning fingerprint orientation fields using continuous restricted Boltzmann machines. In *Proceedings of the 2013 2nd IAPR Asian Conference on Pattern Recognition*. IEEE, 351–355.

[123] Mihir Sahasrabudhe and Anoop M. Namboodiri. 2014. Fingerprint enhancement using unsupervised hierarchical feature learning. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*. ACM, 2.

[124] Ruslan Salakhutdinov and Geoffrey E Hinton. 2009. Deep Boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS'09)*, Vol. 1. 3.

[125] Muhammad Muneeb Saleem and John H. L. Hansen. 2016. A discriminative unsupervised method for speaker recognition using deep learning. In *Proceedings of the IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP'16)*. IEEE, 1–5.

[126] Pouya Samangouei and Rama Chellappa. 2016. Convolutional neural networks for attribute-based active authentication on mobile devices. In *Proceedings of the IEEE 8th International Conference on Biometrics Theory, Applications, and Systems (BTAS'16)*. IEEE, 1–8.

[127] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61 (2015), 85–117.

[128] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[129] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. 2016. Geinet: View-invariant gait recognition using a convolutional neural network. In *Proceedings of the IEEE International Conference on Biometrics (ICB'16)*. IEEE, 1–8.

[130] Pedro Silva, Eduardo Luz, Rafael Baeta, Helio Pedrini, Alexandre Xavier Falcao, and David Menotti. 2015. An approach to iris contact lens detection based on deep image representations. In *Proceedings of the 2015 28th SIBGRAPI Conference on Graphics, Patterns, and Images*. IEEE, 157–164.

[131] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur. 2016. Deep neural network-based speaker embeddings for end-to-end speaker verification. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT'16)*. IEEE, 165–170.

[132] Nisha Srinivas, Harleen Atwal, Derek C. Rose, Gayathri Mahalingam, Karl Ricanek, and David S. Bolme. 2017. Age, gender, and fine-grained ethnicity prediction using convolutional neural networks for the East Asian face dataset. In *Proceedings of the IEEE 12th International Conference on Automatic Face and Gesture Recognition (FG'17)*. IEEE, 953–960.

[133] Themos Stafylakis, Patrick Kenny, Mohammed Senoussaoui, and Pierre Dumouchel. 2012. Preliminary investigation of Boltzmann machine classifiers for speaker recognition. In *Proceedings of the Odyssey Conference*. 109–116.

[134] Hong-Ren Su, Kuang-Yu Chen, Wei Jing Wong, and Shang-Hong Lai. 2017. A deep learning approach towards pore extraction for high-resolution fingerprint recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'17)*. IEEE, 2057–2061.

[135] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014a. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*. 1988–1996.

[136] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014b. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1891–1898.

[137] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2015. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2892–2900.

[138] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2016. Sparsifying neural network connections for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4856–4864.

[139] Ilya Sutskever. 2013. *Training Recurrent Neural Networks*. Ph.D. Dissertation. University of Toronto.

[140] Jan Svoboda, Jonathan Masci, and Michael M Bronstein. 2016. Palmprint recognition via discriminative index learning. In *Proceedings of the IEEE 23rd International Conference on Pattern Recognition (ICPR'16)*. IEEE, 4232–4237.

[141] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[142] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1701–1708.

[143] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2015. Web-scale training for face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2746–2754.

[144] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Representation learning by rotating your faces. arXiv:1705.11136.

[145] J. F. Vargas, M. A. Ferrer, C. M. Travieso, and J. B. Alonso. 2011. Off-line signature verification based on grey level information using texture features. *Pattern Recognition* 44, 2 (2011), 375–385.

[146] J. Francisco Vargas, Miguel A. Ferrer, Carlos M. Travieso, and Jesús B. Alonso. 2007. Off-line handwritten signature GPDS-960 corpus. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR'07)*.

[147] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*. IEEE, 4052–4056.

[148] Vasileios Vasilakakis, Sandro Cumani, and Pietro Laface. 2013. Speaker Recognition by Means of Deep Belief Networks. Retrieved from https://cls.ru.nl/staff/dvleeuwen/btfs-2013/vasilakakis-btfs2013.pdf.

[149] Chen Wang, Junping Zhang, Jian Pu, Xiaoru Yuan, and Liang Wang. 2010. Chrono-gait image: A novel temporal template for gait recognition. In *Proceedings of the European Conference on Computer Vision*. 257–270.

[150] Ruxin Wang, Congying Han, and Tiande Guo. 2016. A novel fingerprint classification method based on deep learning. In *Proceedings of the IEEE 23rd International Conference on Pattern Recognition (ICPR'16)*. IEEE, 931–936.

[151] Xiaolong Wang, Rui Guo, and Chandra Kambhamettu. 2015. Deeply-learned feature for age estimation. In *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 534–541.

[152] Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10, 2 (2009), 207–244.

[153] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*. 499–515.

[154] Lior Wolf, Tal Hassner, and Itay Maoz. 2011. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, 529–534.

[155] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. 2016. Multi-view gait recognition using 3D convolutional neural networks. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'16)*. IEEE, 4165–4169.
[156] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. 2016. A comprehensive study on cross-view gait based human identification with deep CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 2, 209–226.
[157] Xingpeng Xu, Zhenhua Guo, Changjiang Song, and Yafeng Li. 2012. Multispectral palmprint recognition using a quaternion matrix. *Sensors* 12, 4 (2012), 4633–4647.
[158] Takanori Yamada, Longbiao Wang, and Atsuhiko Kai. 2013. Improvement of distant-talking speaker identification using bottleneck features of DNN. In *Proceedings of the INTERSPEECH Conference*. 3661–3664.
[159] Chao Yan, Bailing Zhang, and Frans Coenen. 2015. Multi-attributes gait identification by convolutional neural networks. In *Proceedings of the 2015 8th International Congress on Image and Signal Processing (CISP'15)*. IEEE, 642–647.
[160] TzeWei Yeoh, Hernán E. Aguirre, and Kiyoshi Tanaka. 2016. Clothing-invariant gait recognition using convolutional neural network. In *Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'16)*. IEEE, 1–5.
[161] Dit-Yan Yeung, Hong Chang, Yimin Xiong, Susan George, Ramanujan Kashi, Takashi Matsumoto, and Gerhard Rigoll. 2004. SVC2004: First international signature verification competition. In *Biometric Authentication*. Springer, 16–22.
[162] Dong Yi, Zhen Lei, and Stan Z. Li. 2014. Age estimation by multi-scale convolutional network. In *Proceedings of the Asian Conference on Computer Vision*. 144–158.
[163] Weidong Yin, Yanwei Fu, Leonid Sigal, and Xiangyang Xue. 2017. Semi-Latent GAN: Learning to generate and modify facial images from attributes. arXiv:1704.02166.
[164] Shiqi Yu, Daoliang Tan, and Tieniu Tan. 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 4. IEEE, 441–444.
[165] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*. 818–833.
[166] Cheng Zhang, Wu Liu, Huadong Ma, and Huiyuan Fu. 2016b. Siamese neural network based gait recognition for human identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16)*. IEEE, 2832–2836.
[167] David Zhang, Wai-Kin Kong, Jane You, and Michael Wong. 2003. Online palmprint identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 9 (2003), 1041–1050.
[168] Kaipeng Zhang, Lianzhi Tan, Zhifeng Li, and Yu Qiao. 2016. Gender and smile classification using deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 34–38.
[169] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. 2014. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1637–1644.
[170] Qi Zhang, Haiqing Li, Zhenan Sun, Zhaofeng He, and Tieniu Tan. 2016a. Exploring complementary features for iris recognition on mobile devices. In *Proceedings of the International Conference on Biometrics (ICB'16)*. IEEE, 1–8.
[171] Dandan Zhao, Xin Pan, Xiaoling Luo, and Xiaojing Gao. 2015. Palmprint recognition based on deep learning. In *Proceedings of the 6th International Conference on Wireless, Mobile, and Multi-Media (ICWMMN'15)*. 214–216.
[172] Zijing Zhao and Ajay Kumar. 2017. Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network. *IEEE Transactions on Information Forensics and Security* 12, 5 (2017), 1017–1030.
[173] Yu Zhong, Yunbin Deng, and Anil K. Jain. 2012. Keystroke dynamics for user authentication. In *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'12)*. IEEE, 117–123.
[174] Erjin Zhou, Zhimin Cao, and Qi Yin. 2015. Naive-deep face recognition: Touching the limit of LFW benchmark or not? arXiv:1501.04690.
[175] Linnan Zhu, Keze Wang, Liang Lin, and Lei Zhang. 2016. Learning a lightweight deep convolutional network for joint age and gender recognition. In *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR'16)*. IEEE, 3282–3287.
[176] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2013. Deep learning identity-preserving face space. In *Proceedings of the IEEE International Conference on Computer Vision*. 113–120.

Received November 2016; revised February 2018; accepted February 2018