Arachchilage and Izquierdo *EURASIP Journal on Image and Video Processing* (2020) 2020:25
https://doi.org/10.1186/s13640-020-00510-w

EURASIP Journal on Image and Video Processing

REVIEW Open Access

Deep-learned faces: a survey

Samadhi P. K. Wickrama Arachchilage* and Ebroul Izquierdo

*Correspondence:
s.wickramaarachchilage@qmul.ac.uk
Multimedia and Vision Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Rd, E1 4NS London, UK

Abstract

Deep learning technology has enabled successful modeling of complex facial features when high-quality images are available. Nonetheless, accurate modeling and recognition of human faces in real-world scenarios "on the wild" or under adverse conditions remains an open problem. Consequently, a plethora of novel deep network architectures addressing issues related to low-quality images, varying pose, illumination changes, emotional expressions, etc., have been proposed and studied over the last few years.

This survey presents a comprehensive analysis of the latest developments in the field. A conventional deep face recognition system entails several main components: deep network, optimization loss function, classification algorithm, and train data collection. Aiming at providing a complete and comprehensive study of such complex frameworks, this paper first discusses the evolution of related network architectures. Next, a comparative analysis of loss functions, classification algorithms, and face datasets is given. Then, a comparative study of state-of-the-art face recognition systems is presented. Here, the performance of the systems is discussed using three benchmarking datasets with increasing degrees of complexity. Furthermore, an experimental study was conducted to compare several openly accessible face recognition frameworks in terms of recognition accuracy and speed.

Keywords: Face recognition, Deep learning, Convolutional neural network

1 Introduction

Face conveys a plethora of discriminative features rich enough to determine one's identity [1]. These features can be extracted in unconstrained scenarios and non-intrusive manners. Hence, automated face recognition can be exploited in a large number of practical applications [2]. Among others, it has shown excellent capabilities in security applications like intelligent surveillance [3, 4], user authentication applications like traveler verification at border crossing points [5, 6], and diverse other mobile and social media applications [7–10]. Indeed, person identity prediction based on facial features for practical purposes is a valuable tool in modern information technology [11]. Straightforwardly, as it may seem, the underlying modeling and mapping of faces is complex and it becomes daunting due to the diversity of facial features. Such complexity is further exacerbated by other variations like emotions, illumination, make up, and low-quality sensing [12, 13]. To tackle this important, yet challenging problem of face recognition, intensive research efforts have been reported by numerous research groups and scholars. The discipline can be traced

back to the sixties [14, 15], when both feature based approaches and holistic approaches were reported. Feature-based approaches exploit the geometric relationships among distinctive facial features such as eyes, mouth, and other face landmarks [16–23]. In contrast, holistic approaches aim at capturing features of the entire facial area in an image [24–29]. Holistic approaches assign equal importance to all the pixels rather than special attention to a set of points of interest. Hence, these approaches encompass higher distinctive power at the cost of increased computational complexity [6, 30].

Deep convolutional neural networks (DCNNs) are a holistic approach that recently enabled a quantum leap in the field. In 2014, Facebook reported a face recognition system named DeepFace [27] which achieved near-human performance on LFW benchmark [31]. This accuracy was quickly surpassed by systems like DeepId3 [28] and FaceNet [29]. Such substantial progress of face recognition technology is a reflection of cutting-edge research developments in deep network architectures. Starting from LeNet in 1989 [32], DCNNs have evolved into sophisticated networks particularly fueled by classification challenges like The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [33]. AlexNet [34], VGGNet [35], and GoogleNet [36] are arguably the three most influential ILSVRC networks.

The role of a deep network in a classic image classification system is to map the complex high-dimensional image information into a low-dimensional proprietary template, i.e., feature vector. The generated feature vectors can be interpreted as points in a fixed-dimensional space. Clearly, face images are a subspace in the much larger image space. This fact implies that network architectures that succeeded in the problem of image classification are adaptable to face classification. Some successful face recognition applications that emerged from image classification networks are as follows: DeepId3 [28] which was influenced by VGGNet [35] and GoogLeNet [36], Google's Facenet [29] which used GoogleNet [36] architecture, and VGGFace [37] that exploited concepts from VGGNet [35].

A deep network is generally underpinned by an optimization loss function. When the deep net outputs feature vectors from input images, the loss function adds discriminative power to the generated features. Over the years, loss functions have evolved complementing the network architectures. These loss functions can be categorized as classification based approaches, i.e., softmax loss and it's variants, and metric learning approaches, i.e., contrastive loss and triplet loss. Successful exploitation of suitable loss functions in face recognition includes softmax loss in DeepFace [27], a variation of softmax loss as used in Arcface [38] and a tripletloss used in FaceNet [29].

Figure 1 shows the data flow of a typical face recognition system. During training, the network model learns from large training datasets. The trained model is then used to generate feature vectors for test faces. A classic face recognition task generally includes a gallery of labelled faces and probe/query images. Labelled gallery images are usually processed in advance in a step called 'enrolment process'. Here, the feature vectors/templates of the gallery subjects are generated. These features are then either stored with their corresponding labels or used to generate subject specific models. During the face recognition phase, the template of the query face is compared to the enrolled templates. This comparison can either use a nearest neighbor search or a model based classification. The former approach is referred throughout this paper as *template learning* and the latter is referred as *subject-specific modelling*.

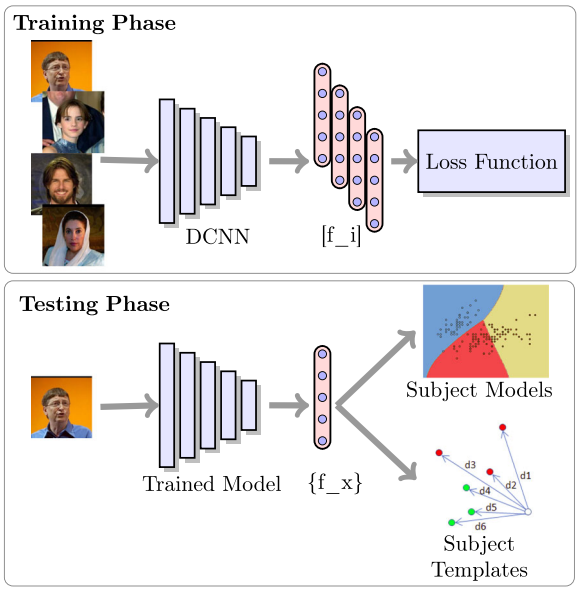

Fig. 1 The dataflow of a deep face recognition system. During training the network learns feature representations (f_i) of faces by getting gradually penalized by the loss function. During testing, the pre-trained model is used to generate features of test faces. The generated features are classified/compared for identity determination

An important aspect of face recognition is benchmarking. As mentioned before, network architectures together with optimization loss functions and sufficient and diversified train datasets have enabled successful modeling of complex facial features when provided with high-quality images. These face recognition systems reported near-perfect performance on classic benchmarks like LFW [31]. However, the performance saturation on these benchmarks resulted in more challenging benchmarks [39–41] entailing more realistic pictures captured under adverse conditions. The evaluations on such real-world data shows that the performance of face recognition systems is affected by many factors including emotions, illuminations variations, make up and pose variations [39, 40, 42, 43].

1.1 Surveys on deep face recognition

Due to the importance of the topic and the vast number of face recognition papers reported in the past, there is indeed no shortage of related surveys either. Some noteworthy face recognition surveys include Zhang et al. [11], Jafri et al. [30], Bowyer et al. [44], and Scheenstra et al. [45]. These comprehensively survey face recognition systems prior to DeepFace. Hence, these surveys do not discuss the new sophisticated deep learning approaches that emerged during the last decade. Surveys that discuss deep face recognition have singled out face recognition as an individual discipline rather than a collection of components adopted from different studies. These surveys generally discuss the face recognition pipeline: face pre-processing, network, loss function, and face classification [42, 46, 50] or discuss a single aspect of face recognition such as 3-D face recognition [47], illumination face recognition [52] or pose invariant face recognition [51]. Although these surveys are important and provide an excellent basis for the analysis of the state-of-the-art in the field, they do not provide conclusive comparisons or analysis of the underlying network architectures.

To better illustrate the difference of the key contributions in the past and this survey, Table 1 summarises the main deep face recognition surveys. The analysis presented by Wang et al. [46] is arguably the most comprehensive survey yet in the field. It provides a holistic overview of the broad topics of deep face recognition including the face recognition pipeline, face datasets, benchmarks, and industry scenes, briefly surveying all elements of face recognition. In contrast, this paper focuses on deep learning based components in the recognition pipeline and delivers a much detailed analysis of the 18 most critical deep face recognition systems. The paper describes a face recognition system as a unique combination of a deep net, loss function, classification approach, train dataset, and other system specific novelties if any. To properly understand how each system was derived, the paper also discusses the evolution of the aforementioned components.

1.2 Paper contribution

The key contributions of this survey include:

Table 1 Surveys that discuss deep face recognition

Year	Survey	Contribution
2019	Deep-learned faces: a survey (Ours)	Provides an elaborated study on 18 state-of-the-art face recognition systems. Discusses the origin and evolution of these systems providing insights of how the network designs and algorithms were derived from image recognition and adapted to face recognition. The systems are analyzed based on reported performance and experimental results, comparing the performance against dataset quality and complexity.
2019	Deep face recognition: a survey [46]	Provides an overview of all the topics of deep face recognition covering algorithm designs, databases and protocols, application scenes, reported benchmark results, etc. A good reference for a quick, shallow summary across the broad discipline.
2018	Deep face recognition: a survey [42]	Summarizes the advances of deep face recognition techniques from 2014–2018. Includes a summary on face datasets, face pre-processing, alignment, network architectures and loss functions. Additionally, discusses the performance of the face recognition systems with respect to the IJB-A [39] dataset.
2018	3D face recognition: a survey [47]	Categorizes 3D face recognition systems into pose-invariant recognition, expression-invariant recognition and occlusion-invariant recognition. Provides an overview of publicly available 3D face databases.
2018	A comprehensive analysis of LBCNN for fast face recognition in surveillance video [48]	Comparatively analyses Local Binary Convolutional Neural Networks (LBCNNs) against other state-of-the-art networks in terms of sensibility and processing time.
2017	A survey on facial feature extraction techniques for automatic face annotation [49]	Discusses six facial feature extraction approaches: speeded up robust features, eigenfaces, scale invariant feature transform, convolutional neural network, gabor filter, and local binary pattern.
2016	A survey of deep face recognition in the wild [50]	Discusses the network models of seven face recognition systems, comparing their reported performance on LFW [31] benchmark.
2016	A comprehensive survey on pose-invariant face recognition [51]	Discusses pose-robust facial feature extraction systems under two categories: engineered features and learning based features. Deep frameworks are discussed under learning based features.
2015	Addressing the illumination challenge in two-dimensional face recognition: a survey [52]	Provides summarized review of 72 state-of-the-art illumination-invariant facial feature extraction methods prior to 2014.
2015	A survey of unconstrained face recognition algorithm and its applications [53]	Discusses face recognition techniques in terms of their behavior at pose variations, non-uniform motion blur and illumination for the period 2011–2014. The discussion includes several neural network based systems.

Deep learning based face recognition is either the main focus or is included as a subsection in each survey

Table 2 DCNN frameworks that had significant impact on face recognition

Year	Publication	Contribution	Network architecture	ILSVRC error (top-5) (%)	Face recognition system
2012	AlexNet [34]	Large DCNN with 60M parameters and 650,000 neurons	Ensemble of 7 models	15.3	-
2014	VGGNet [35]	Increasing depth using very small convolution filters.	Ensemble of 2 models	6.8	VGGFace [37]
					DeepId3 [28]
2014	GoogleNet [36]	Inception architecture	Ensemble of 7 models	6.67	FaceNet [29]
					DeepId3 [28]
					OpenFace [65]
2014	InceptionV2[73]	Adding batch normalization to Inception architecture	Batch Normalized Inception ensemble	4.9	-
2014	InceptionV3 [74]	Adding factorization to Inception architecture	Ensemble of four Inception-V3s	3.58	-
2015	Residual Learning [70]	A residual learning framework to ease the training of very deep CNNs	ResNet 34	5.60	DLIB [64]
			ResNet 50	5.25	ArcFace [38]
			ResNet 101	4.60	CosFace [61]
			ResNet 152	4.49	SphereFace [60]
			Ensemble	3.57	-
2016	InceptionV4 [71]	Adding residual learning on top of Inception	Inception-ResNet-v1	4.3	FaceNet_Re [66]
			Inception-ResNet-v2	3.7	of
			Ensemble of 4 DNNs	3.1	

- The background knowledge required to understand and analyze the underlying frameworks used in face recognition, including,
  - The origin and evolution of DCNN frameworks that were effective in face recognition (Table 2).
  - The loss functions used in face recognition, categorized and compared under two classes: classification based approaches and metric learning approaches.
  - A comparative discussion on two main classification approaches used in face recognition, i.e, template learning and subject specific modelling.
  - A brief discussion on key face datasets and evaluation benchmarks.
- An elaborated discussion on 18 state-of-the-art face recognition systems (DeepFace [27], DeepId [54], DeepID2 [55], DeepID2+ [56], VGGFace [37], DeepID3 [28], FaceNet [29], Baidu [57], NAN [58], Template Adaptation [59], SphereFace [60], CosFace [61], ArcFace [38], B-CNN [62], DCNNmanual+metric [63], DLIB [64], OpenFace [65], and FaceNet_Re [66]).
- The face recognition systems are analyzed based on the network architecture, loss function, classification approach, and train data and other unique system design details.
- The performance of face recognition is discussed based on three scenarios:

  - The performance on good quality data (LFW [31] benchmark)
  - The performance on unconstrained data (IJB-A [39] benchmark)
  - The performance under millions of gallery distractors (MegaFace [67] benchmark)
- An experimental study that compares three face recognition systems (DLIB [64], OpenFace [65], and FaceNet_Re [66]) with respect to face recognition accuracy and speed.
- Discussion on open issues and challenges in face recognition highlighting possible future research.

The remainder of the survey is organized as follows. Section 2 presents a cognitive study of the evolution of DCNN architectures. Then, the paper presents a comparative analysis of loss functions in Section 3, a study of classification algorithms in Section 4, and face datasets and evaluation benchmarks in Section 5. Section 6 presents a study on state-of-the-art face recognition systems. This study is three fold and includes an individual systems analysis, a comparative performance analysis on three benchmarks and an experimental performance analysis. Finally, the paper presents the open issues of face recognition followed by the conclusion.

2 The evolution of deep face architectures

Andrew Ng, the Chief Scientist at Baidu Research, described the notion of deep learning as "Using brain simulations, hope to make learning algorithms much better and easier to use and make revolutionary advances in machine learning and AI". While deep neural networks (DNNs) have conquered different disciplines, convolutional neural networks (ConvNets or CNNs) have been particularly effective in visual science [68]. Given the appropriate network architecture, CNNs are able to process, analyze, and classify high-dimensional patterns, resulting in an extremely valuable tool in computer vision.

A typical DCNN adheres to a conventional structure which consists of a set of stacked convolutional layers followed by contrast normalization and max-pooling and finally one or more fully connected layers [36]. Different variants of this structure have been explored for performance enhancements. Please refer to Fig. 5 for the general structure of a DCNN.

The evolution of deep network architectures initiated with increased size with respect to depth, the number of levels, and width, the number of units at each level [34, 35, 69]. Nonetheless, the increased complexity associated with larger nets was not favored in practical applications. Hence, systems like GoogleNet pioneered architecturally enhanced networks with lesser parameters [36]. This was followed by Microsoft's efforts to simplify the training process by using networks with lesser complexity [70]. In the immediate history, researchers have combined these two design techniques for further simplified networks [71].

Classification challenges such as ILSVRC [33], MNIST, and CIFAR have led to several milestone in image recognition. AlexNet [34], the winner of ILSVRC 2012, achieved a top-5 test error rate of 15.3%, which is the pioneer of DNN-based image recognition. To this day, the publication is considered to be one of the most influential breakthroughs. The second milestone was recorded when VGGNet [35], the second place winner ILSVRC 2014, achieved significant improvements (top 5 test error rate of 6.8%) with increased depth in DNNs.

Despite the fact that going deeper with convolutions seemed to be the straightforward solution for accuracy enhancements [34, 35, 69], this approach had two main drawbacks: (1) the large number of parameters that these deeper networks encompassed made the network prone to over-fitting and (2) the deeper networks meant increased computer resource consumption. These factors turned the attention of research community towards sparsely connected systems. Nonetheless, sparse systems were not a simple solution and possessed complications and limitations. The calculations associated with these non-uniform sparse systems, even if the number of arithmetic operations were fewer, suffered from the overhead of look-ups and cache misses. In comparison, dense nets, even with higher number of arithmetic operations, had the advantage of fast dense matrix multiplication operations provided by improved numerical libraries [34, 72].

GoogleNet [36], which was the winner of ILSVRC 2014, introduced an architecture code-named Inception, which was capable of outperforming AlexNet with 12 times fewer parameters as that of AlexNet. The main concept behind this architecture is finding optimal local sparse structure covered by readily available dense components.

The inception architecture was learned layer by layer. In a single layer, units with high correlation were clustered together. These clusters which are connected to the layer units were considered as the next layer. When these inception modules were stacked, the higher layers required more and more 3×3 and 5×5 convolutions. This is because the highly abstract features are captured by the higher layers, and their spatial concentration reduces as a result. To avoid such complexities, dimension reduction was introduced to the architecture. In doing so, 1×1 convolutions were introduced before the 3×3 and 5×5 convolutions so that these 1×1 convolutions can compute reductions prior to feeding them to more expensive convolutions. The Inception architecture was later modified in the subsequent versions by adding batch normalization in Inception V2 [73] and additional factorizations in Inception V3 [74].

The uniqueness of Inception architecture is that the design principles focus more on computational simplicity, enabling the inference to be run even on a single machine. Due to this nature of GoogleNet, it was later used by many face recognition systems including Google's FaceNet [29] and DeepId3 [28].

In a contemporary research, the Microsoft Research employed the concept of deep residual learning [70] for image recognition. The authors show that the residual learning framework enables very deep networks, deeper than the traditional DNNs, to be implemented with lesser complexity. The study presented a DNN with 152 layers, which is eight times as deep as VGGNet [35].

Residual learning can be explained as follows. Consider a set of layers stacked, this could be the entire network or a part of it. Let the input to the stack of layers be x and the underlying mapping be $H(x)$. Instead of training the layers to learn the traditional complicated function, the stack of layers are trained to learn the corresponding residual function, i.e., $H(x) - x$ thus deriving Eq. 1.

$$\begin{aligned} F(x) &= H(x) - x \\ F(x) + x &= H(x) \end{aligned} \tag{1}$$

The authors presented several networks of different sizes. ResNet-152, which is 152 layers deep, outperformed VGGNet and GoogleNet in ImageNet validation with a top

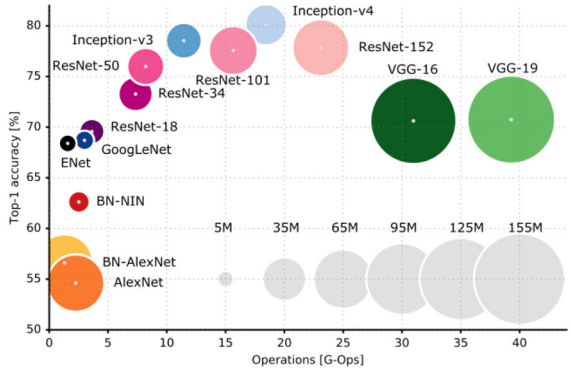

Fig. 2 Top 1 one-crop accuracy (using only the center crop) on the ImageNet-1k validation set with respect to the computational complexity of the model. The computational complexity is measured as the number of floating-point operations (FLOPs) required for a single forward pass. The size of each ball corresponds to the number of learnable parameters

5 error of 4.49%. An ensemble of ResNets which achieves 3.57% top 5 error won the ILSVRC-15.

The residual networks, even though much deeper, have lower complexity than the traditional DNNs. An architectural comparison between VGG-19 model, a 34-layer deep network and the same 34-layer network with residual connections (ResNet-34), explains the complexity reduction in ResNets. VGG-19 model has 19.6 billion FLOPS whereas the 34-layer deep networks, both plain and with residual connections, have only 3.6 million FLOPS each. The plain net and ResNet both have the same FLOPs because the identity mappings do not introduce any parameters nor computational complexity. Despite the lesser complexity, ResNet34 outperformed the VGG-19 model in ImageNet validation.

When residual connections on top of a traditional DCNN architectures achieved closer performance to that of Inception V3, it raised the question whether residual connections on top of Inception would further enhance the performance. This hypothesis was explored in Inception V4 (Fig. 3) [71]. The authors showed that, while it is feasible to achieve competitive results through very deep networks without the use of residual connections, inclusion of residual connections in fact improves training speed in a greater scale.

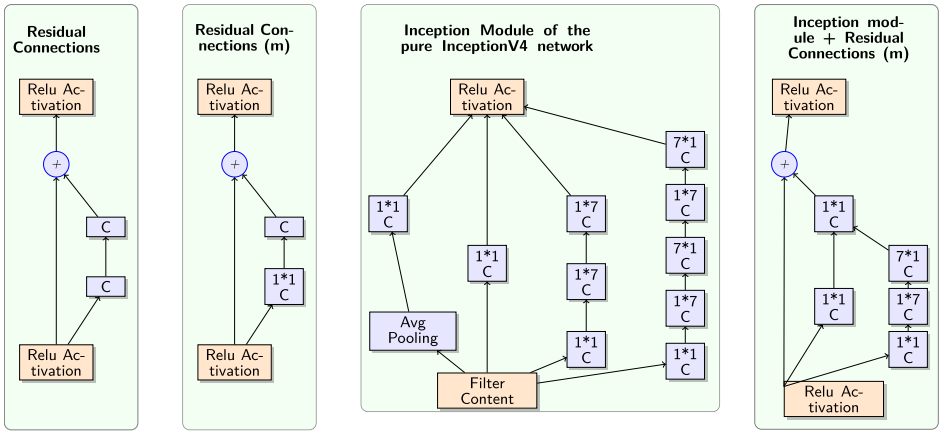

Fig. 3 From left to right in order are original residual connections [70], modified residual connections used in [36], the schema for 17×17 grid of the pure InceptionV4 network, and the schema for 17×17 grid (Inception-ResNet-B) module of Inception-ResNet-v1 network

In addition to discussed networks, bilinear CNNs are a model designed for image recognition and later adopted in face recognition. The network consists of two feature extractors whose outputs are multiplied using outer product at each location of the image and pooled to obtain a bilinear vector [75]. This model was proven to be effective in fine-grained recognition tasks.

The major architectural innovations in DCNN history are associated with three concepts: increased network size, inception architecture, and residual connections. These innovations vary in performance indices like model complexity, computational complexity, memory usage, and inference time. These indices are vital in selecting an appropriate architecture compatible with the resource constraints in practical deployment. Canziani et al. [76] and Bianco et al. [77] presents an experimental comparison between different DNNs. From the results of Canziani et al., Fig. 2 presents the model complexity and computational complexity of DCNNs that have major impact on face recognition (with the exception of E-Net, BN-NIN, and BN-AlexNet).

3 Comparative analysis of loss functions

The loss function is the supervisory signal used to train a deep network. The study of loss functions has been carried along two main lines of research: Fig. 4 (1) classification-based approaches (conventional softmax classifier [27, 37, 78] and modified versions of softmax loss [38, 60, 61, 79–86]) and (2) metric learning approaches (contrastive loss [28, 55, 56, 87, 88] and triplet loss [29]). The softmax loss learns by classifying each train image into one of the pre-defined classes. Variants of softmax loss have made efforts to increase the intra-class compactness in the process. In contrast, metric learning approaches learn by increasing the similarity between faces of same identity while decreasing the similarity between the faces of different identities. Regardless of the approach, all deep face supervisory signals are driven towards a single goal, inter-class discrepancy with intra-class compactness.

3.0.1 Classification-based approaches

The softmax loss is a multi-class classification problem where the input data contains one or more images of a set of individuals and the classifier learns the features of each individual. Despite being referred as softmax loss for convenience, technically, a k-way softmax function is employed to obtain a probability distribution over labels of k classes [27]. And the minimization is carried out for the cross-entropy loss for each training sample. The softmax loss is denoted in Eq. 2,

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{2}$$

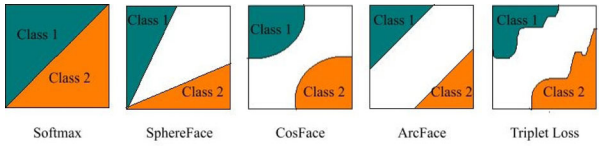

Fig. 4 Classification boundaries of different loss functions in binary classification

where $x_i \epsilon R^d$ denotes the d-dimensional deep feature of the ith sample, belonging to the y_i^{th} class. $W_j \epsilon R^d$ denotes the jth column of the weight $W \epsilon R^{d*n}$ and $b_j \epsilon R^n$ is the bias term. The batch size and the class number are N and n, respectively.

Softmax loss, despite achieving inter-class dispersion, provides no particular inclination towards intra-class compactness. Hence, the features learned through softmax loss may not be discriminative enough for rather challenging open-set classification problem [38]. Studies that followed have reported several efforts to enhance the discriminative power of softmax loss [60, 61, 79–81, 83–86]. An extension of softmax loss named center-loss [79] attempted to achieve the missing intra-class compactness by taking into account the euclidean distance between the feature vector and the center of the class. The authors show that a combination of center-loss and the softmax loss could be an optimum solution. However, in the matter of huge training datasets with a large number of classes, the class-wise learning approach becomes complicated and difficult. In an effort to solidify class-wise learning in large datasets, a new approach named SphereFace [60] incorporated a multiplicative angular margin penalty. Even though a new loss function was introduced in this publication, the presented optimum solution was a hybrid with softmax loss. Later, Wang et al. [61] proposed a system named CosFace which used a cosine margin penalty. As opposed to Sphereface, CosFace was an additive margin. This approach outperformed Sphereface.

Most recently in 2019, a research team from Imperial college introduced an additive angular margin loss named ArcFace [38]. The derivation of ArcFace can be outlined as follows.

Consider the traditional softmax loss denoted in Eq. 2. The bias is removed and the logit $W_j^T x_i$ is transformed to its dot product as $W_j^T x_i = \| W_j \| \| x_i \| \cos\theta_j$. When l_2 normalization is applied on individual weight and embedding feature , $\| W_j \| = 1$ and $\| x_i \|$ is re-scaled to s yielding the following equation.

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{scos\theta_{y_i}}}{e^{scos\theta_{y_i}} + \sum_{j=1,j!=y_i}^{n} e^{scos\theta_j}}$$

Now, the predictions only depend on the angle between the feature and the weight. The inter-class discrepancy and intra-class compactness is achieved by the additive angular margin penalty m; hence, the final equation is as follows.

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{scos\theta_{y_i}+m}}{e^{scos\theta_{y_i}+m} + \sum_{j=1,j!=y_i}^{n} e^{scos\theta_j}}$$

ArcFace system reported considerable improvement reporting 99.83% LFW accuracy.

3.0.2 Metric learning approaches

Metric learning approaches are a different optimization approach than softmax loss or its variants. In metric learning, the network is provided with sample images and is penalized based on whether the samples are of the same class or not. Contrastive loss and triplet loss are two metric learning approaches popular in face recognition.

Contrastive loss is generally used in Siamese style networks. A Siamese network is an architecture with two parallel neural networks with shared weights. Each network takes a different input, and the two outputs are combined to provide some prediction [89]. Contrastive loss was proposed by Hadsell et al. [90] and was used in face recognition

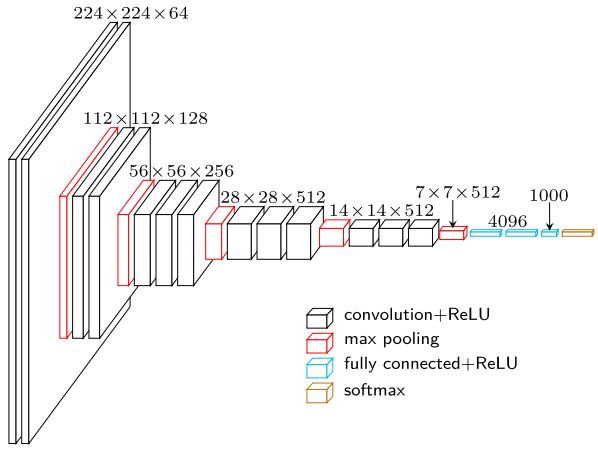

Fig. 5 Architecture of VGG16 model

systems like DeepId2 [55], DeepId2+ [56], DeepId3 [28], and others [87, 88]. Figure 7 shows the Siamese network used in DeepId2+.

The researchers of Google presented a system that learns a direct mapping from face images to discrete points in the compact euclidean space [29]. The optimization loss function is triplet loss. Given a triplet (an anchor, positive sample and negative sample), this loss aims at minimizing the distance between the anchor and its positive while maximizing the distance between the anchor and the negative. Contemporary research carried out by Baidu Research also reports the use of triplet loss [57].

Despite being conceptually straightforward, the effectiveness of metric learning mainly depends on the input samples. For example, FaceNet uses a hard sample mining algorithm for optimum triplets. Moreover, the number of possible triplets grows exponentially with dataset size and hence effective triplet mining becomes complicated. Studies that followed [37, 38] reported that while triplet loss is an effective approach, learning by classification and metric learning approaches makes the training more convenient.

4 A study on face classification algorithms

Generally, the train data in face recognition are large scale datasets diversified with variations in gender, ethnicity, profession, etc. In contrast, gallery set is much smaller and application specific (e.g., mugshot images of persons of interest). Often times, gallery images are disjoint from the train data. Even if the gallery set was included in the much larger train set, each update to gallery will require complete retraining of the network. In these situations, the effort to use the trained model without alteration, for online face recognition, is inconvenient and naive. To this end, deep face recognition exploits the strategy of transfer learning. In this approach, as shown in Fig. 6, the network learns from large volumes of train data and the trained model is used to generate features for test faces. A shallow classifier is then used on the generated features for face identification. In doing so, the enrolment of gallery samples, i.e., training the shallow classifier, is carried out as an intermediary step between offline model training and online face recognition. The enrolment could be a model based approach or a template learning approach. This section aims to discuss the algorithms used in this shallow classification process.

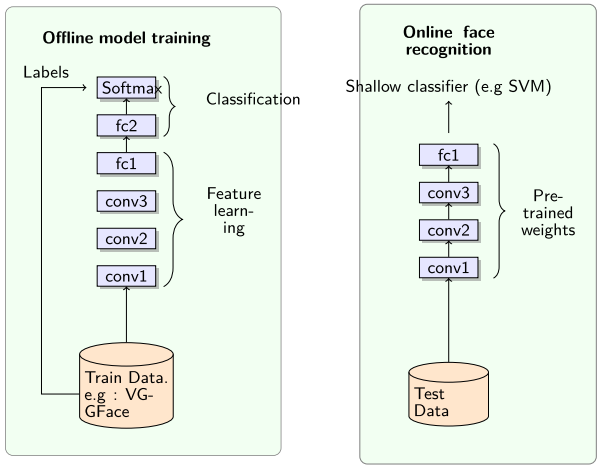

Fig. 6 Feature based transfer learning with pre-trained models. conv, convolution layer; fc, fully connected layer

Generally, transfer learning includes a source domain which is trained offline and a target domain for online processing [59, 91, 92]. In this context, the source domain is the large datasets used for offline training of the network model and the target domain is the online face recognition data. Prior to DeepFace, transfer learning meant fine-tuning the network model with the gallery samples. DeepFace presented a varied approach of transfer learning for face recognition. The DeepFace net was initially trained as a traditional multi-class face classification problem. The authors considered the output of the last fully connected layer as a raw feature representation of the input face. With this notion, DeepFace used two identical DNNs with shared weights to simultaneously generate feature vectors for two faces for face verification. A contemporary research that exploited a similar concept is the DeepId series [28, 54–56], which used the generated feature vectors for tasks like face verification and recognition. This feature vector based classification was exploited in face recognition in two main approaches: (1) subject specific modelling and (2) template based learning.

When several gallery images are available for a single subject, it results in multiple feature vectors per subject. These feature vectors can be modeled into a single representation for the subject. This is generally carried out with the use of algorithms like support vector machines (SVM). The model-based approaches yield optimum performance multiple imagery per subject is available. In other circumstances, template-based learning is a straightforward approach. In template learning, the unknown feature vector, i.e., template, is compared to known templates to calculate the nearest neighbor.

In contrast to image-based face recognition, video-based face recognition generally has more than one face image for a probe subject spread across a set of consecutive frames. Hence, multiple feature vectors are available for comparison [58, 62, 63, 93, 94] during the classification. The studies on video face recognition has carried out classification in three main approaches: (1) perform classification on each frame, (2) result pooling over set of frames [62, 63, 94], and (3) integration of information across frames for one-time face recognition [58, 95]. The second and third approaches maintain the information across all frames and have reported progress on IJB benchmark.

5 Face datasets and evaluation benchmarks

The data serves two purposes in a typical face recognition system; it serves as training data and as benchmarks for system validation. It known that the quality of train data has a huge impact on the performance of a DNN. Similarly, the quality of the validation data has a huge impact on the reliability of the benchmark results. The term "quality" refers to the size and the level of inter and intra-class variations. The intra-class variation is a measure of the depth of the dataset, i.e., the number of images per each individual and the inter-class variations is achieved by increasing the breadth, the number of individuals in the dataset (Table 3).

Initially, face datasets consisted of high-quality images mostly featuring celebrities [31, 88, 96]. Datasets that followed were more practicality driven and hence consisted of data captured at unconstrained environments (e.g., surveillance footages) [39, 40]. Moreover, several datasets aimed at including challenging variations like age gaps [97–100], pose [101] , disguise [102], and ethnic variations [37, 67].

Over the years, face recognition systems have been employing train datasets of increasing scale. Facebook once used a dataset of 500 million images of over 10 million subjects for training face recognition models [103] and Google used a dataset of 4 million facial images from 4000 subjects [27]. The success of these systems, backed up by large-scale private datasets, attracted research attention towards large and openly accessible face datasets like VGGFace2 [78].

The evaluation benchmarks are generally disjoint from the train datasets. They provide an estimate on the reliability of the trained model under different protocols like face verification, closed-set face identification, and open-set face identification [104]. For an unbiased comparison, the results are denoted in notations specified by the benchmark.

Table 3 The benchmark datasets and train datasets for face recognition

Year	Dataset	Media	Subjects
Benchmarks			
2007	LFW [31]	13,233 Images	5749
2011	YTF [96]	3425 Videos	1595
2015	IJB-A [39]	5712 Images 2085 Videos	500
2016	MegaFace [67]	4.7M Images	690,572
2017	IJB-B [40]	11,754 Images 7011 Videos	1845
2018	IJB-C [41]	31,334 Images 11,779 Videos	3531
Train Data			
2013	CelebFaces [106]	87,628 Images	5436
2014	CASIA-WebFace [88]	494,414 Images	10,575
2014	Google (P) [27]	4.4M Images	4K
2015	Facebook (P)[103]	> 500M Images	> 10M
2015	Baidu (P)[57]	1.2M	18K
2015	VGGFace [37]	2.6M Images	2622
2016	MS-Celeb-1M [107]	10M Images	100K
2018	VGGFace2 [78]	3.31M Images	9131
2018	CosFace (P) [61]	5M Images	> 90K

P private datasets

Face verification is the task of determining if two faces belong to the same identity or not. Verification accuracy is generally represented in the receiver operating characteristic (ROC) [31]. The curve plots variance of the true acceptance rate against the false acceptance rate. Closed-set face identification is the task of identifying a probe against the pre-defined gallery with the assumption that the probe has a mate in the gallery. The accuracy of closed-set recognition is commonly denoted using cumulative match characteristic (CMC) [39, 40]. The CMC curve measures the percentage of true identifications within a given rank, i.e., rank 5 identification accuracy denotes the true identifications within the top 5 predictions. Open-set face identification is the task of identifying a probe against the pre-defined gallery while being open to the possibility that the probe may not have a mate in the gallery. The open-set face recognition accuracy can be denoted using decision error trade-off (DET) [39]. The DET curve to plots the false-negative identification rate (FNIR) as a function of false-positive identification rate (FPIR).

This sections aims to provide an overview of face datasets that have been effective in face recognition discussing their important features, advantages, and disadvantages.

5.0.3 LFW [31]

LFW is by far the most effective benchmark for unconstrained face recognition. The dataset comprises 13,233 images of 5749 people under varying conditions of pose, lighting, focus, resolution, etc. The cropped faces are detections of Haar cascade-based face detector by Viola and Jones [105].

The benchmark targets the pair matching problem/face verification. Two evaluation protocols are provided: (1) restricted, the pairs are provided, and (2) unrestricted, the pairs can be generated as per user's preference. The ROC curve is used for recording the results.

5.0.4 YTF [96]

Following LFW, a similar dataset and a benchmark was released with the purpose of evaluation of face recognition in videos under unconstrained category. The dataset comprises 3425 videos of 1595 individuals. These individuals are a subset of those of the LFW dataset.

Since the dataset was designed so as to align with LFW, the benchmark tests were designed the same way. The benchmark includes pair matching tests under two protocols restricted and unrestricted.

5.0.5 VGGFace [37]

VGGFace [37] consists of 2.6 million images of 2 622 individuals. Despite being recognized as one of the largest publicly available datasets for training, the refined dataset where label noise is removed by human annotators, consisting of 800, 000 images.

5.0.6 VGGFace2 [78]

VGGFace2 consists of 3.31 million images of 9131 s classes giving an average of 362.6 images per class. The dataset was created with the aim of achieving a higher depth and breadth. The additional design goals of the dataset include achieving wide range of age, pose, and ethnic variations.

5.0.7 CASIA-Webface [88]

The CASIA-Webface dataset which consists of total of 453,453 images over 10,575 identities. The data is collected from IMDb website. The dataset is designed to be compatible with LFW benchmark, meaning that there are no any overlappings between the two datasets. Hence, a system trained on CASIA-Webface can be independently evaluated on LFW.

5.0.8 CelebFaces [106]

CelebFaces contains 87,628 face images of 5436 celebrities from the Internet, with approximately 16 images per person on average.

5.0.9 Ms-celeb-1m [107]

Ms-celeb-1m dataset consists of a benchmark test which includes evaluation data and evaluation protocol and a separate dataset for training. The evaluation dataset comprises data from one million celebrities and the training dataset comprises approximately 10 million images of 100,000 celebrities.

5.0.10 MegaFace [67]

MegaFace challenge evaluates the performance of face recognition and face verification with up to 1 million distractors. Moreover, it includes protocols for age invariant face recognition. The probe data collection of MegaFace is composed of two datasets: (1) FaceScrub dataset [108] which consists of 100,000 photos of 530 celebrities and (2) FG-Net dataset [109, 110] which consists of 975 photos of 82 people. The latter encompasses variations of age with photos spanning many ages of each subject. The MegaFace distraction data, i.e., gallery collection, includes 1 million photos of more than 690,000 unique subjects collected from Yahoo's Flickr dataset [111].

The evaluation protocol for face recognition is as follows. Let the probe set have M faces of a subject, out of which one is placed in the gallery of 1 million distractors. The face recognition system is provided with the remaining M-1 images. The system is expected to learn from these M-1 images and rank the distractor set in the order of similarity. Ideally, the one image from the probe set should be ranked in the first place. The results are provided via CMC curves. For evaluations on face verification, all pairs between the probe set and distractor set are provided within the dataset. This contains 4 billion negative pairs. The verification results are provided via ROC curves.

5.0.11 IJB [39–41]

In contrast to LFW benchmark which used a commodity face detector, IJB dataset provides a set of face images that are manually aligned (Fig. 8). The manual alignment process aims at preserving challenging variations such as pose, occlusion, and illumination, that are generally filtered out with automated detection. The dataset is a collection of media in the wild which contains both images and videos. The dataset contains media from 500 individuals gathered so as to produce a near-uniform geographic distribution. The complete dataset comprises 5712 images and 2085 videos.

This dataset is benchmarked for face verification and closed-set and open-set face recognition. The performance evaluation on IJB is a process of 10-fold cross validation. The dataset is split 10 random train and test splits with 333 subjects allocated for training at each level and the remaining 167 subjects for testing. The train set can be used

to either fine-tune the network or experimentally derive the optimum threshold distance between two facial feature vectors, which, when exceeded, it can be concluded that the faces are of different identities. The test set is then split into two parts, gallery set and probe set. Each subject has media in both the sets. The media in the probe set are used as the search term and the gallery set is the database that the probe image is tested against. To facilitate open-set classification problem, 55 randomly picked subjects are removed from the gallery. In the protocol specified for face verification, the actual and imposter pairs are provided similar following the LFW convention, but to increase the difficulty, the imposter pairs are selected with restrictions so as to pick pairs of more similarity. The performance is reported using ROC, CMC, and DET curves.

6 State-of-the-art face recognition systems

The conventional face recognition pipeline begins with row input images and followed by pre-processing [112–114], face and facial landmark detection [105, 115–118], alignment [119–124], feature generation, and classification. Although each step along the pipeline has been subjected to research, this survey focuses on the steps controlled by deep learning, i.e., feature generation and classification.

6.1 Study 1 : Individual analysis of system designs

6.1.1 DeepFace [27]

DeepFace uses a nine-layer deep neural network with more than 120 million parameters for face recognition. Softmax loss was employed to train the network, and the train dataset was a private dataset of four million facial images of more than 4000 identities. The system also implements an effective pre-processing mechanism where a 3D model is used to align faces into a canonical pose. In summary, the success of DeepFace is due to three main factors: (1) sound pre-processing step, (2) network architecture, and (3) large scale train data.

In addition to the proposed system, DeepFace also presents an end-to-end face verification system using a Siamese network. Following the training, the network excluding the classification layer is replicated twice to generate features simultaneously for two images. The generated feature vectors are compared in deciding if the two images are of the same person.

6.1.2 DeepId series [28, 54–56]

DeepId introduced the concept that when a CNN is trained for face classification with approximately 10,000 identities and the network is designed such that the number of neurons is reduced as we go higher in the feature extraction hierarchy, it results in the top layers producing compact identity related features with only a few neurons. These identity features, referred to as DeepIds, can then be generalized to other tasks like face verification. This approach of learning facial feature representations through a classification tasks has conceptual similarities to the Siamese network proposed by DeepFace.

The network used in DeepId consists of four convolutional layers, each followed by a max pooling layer. On top of this lies the fully connected layer which is referred to as DeepId layer. The layer was named so because the DeepIds are extracted from this layer. DeepId layer is then followed by the top layer which is a softmax layer. The DeepIds extracted from this network is fed to joint Bayesian technique via which the verification

is carried out. The system was trained on an extended version of CelebFaces [106], codenamed CelebFaces+, which contains 202,599 face images of 10,177 celebrities. The system yielded 97.45% verification accuracy on unconstrained face verification in LFW.

Following the success of DeepId, DeepId2 suggested that including both face identification signals and face verification signals (contrastive loss) for supervision can further increase the accuracy of face recognition/verification systems. This hypothesis was based on the premise that the face identification signals contribute in increasing inter-personal variations whereas face verification signals contribute in reducing intra-personal variations. DeepId2 achieved 99.15% LFW accuracy. This performance was further improved by DeepId2+ which introduced two system improvements: (1) increasing the dimension of hidden representations and (2) introducing supervisory signals to early convolution layers. Please refer to Fig. 7 for DeepId2+ network.

Adding to the continuous improvements, DeepId3 used both identification and verification signals as supervision but on deeper architectures than those of previous DeepId versions. The DeepId3 nets were influenced by VGGNet (stacking of convolutions to achieved increased depth) and GoogLeNet (Inception) architectures. By this implementation, an ensemble of two DeepId3 nets achieved 99.53% LFW accuracy (Fig. 8).

6.1.3 VGGFace [37]

Inspired by VGGNet which showed that deeper convolutions can be more effective in large-scale image recognition, VGGFace applies the same concept for face recognition. The authors employed a modified version of the architecture presented in VGGNet and trained on VGGFace dataset. The authors evaluated two loss functions, softmax loss and triplet loss, and concluded that the triplet loss certainly does provide a better overall performance. Nonetheless, the authors report that training the network as a classifier with softmax loss makes the training significantly easier and faster.

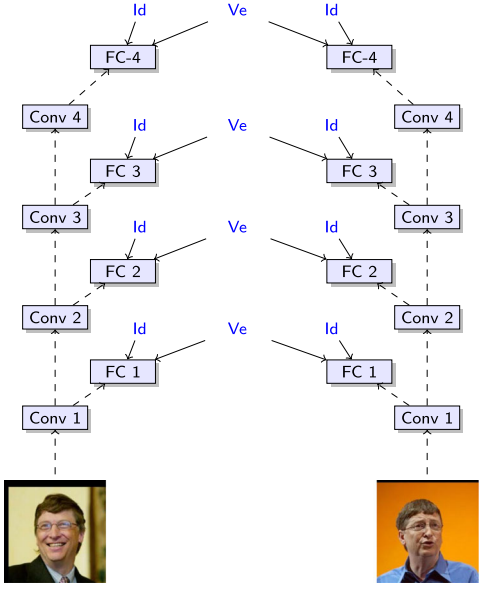

Fig. 7 DeepID2+ net. Conv n, *n*th convolutional layer (with max-pooling). FC n, *n*th fully connected layer. Id, identification supervisory signal. Ve, verification supervisory signal. Dashed arrows denote forward propagation. Solid arrows denote supervisory signals. Nets in the left and right are two DeepID2+ nets with shared weights and different input faces

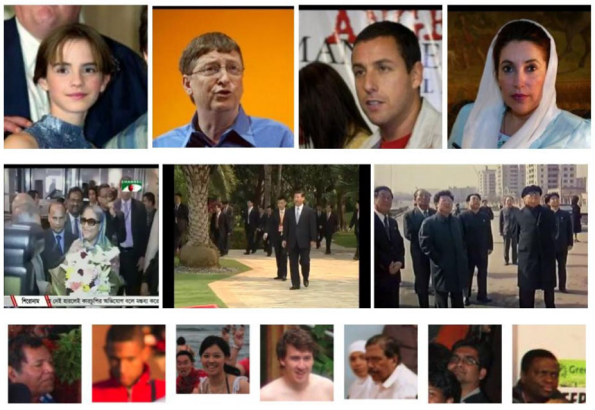

Fig. 8 Sample data in three benchmarks LFW (top), video frames from IJB-A (middle) and MegaFace (bottom)

6.1.4 Template adaptation [59]

The VGGFace system was later used for transfer learning with template adaptation. In this implementation, the deep CNN features from pre-trained VGGNet is combined with linear SVMs trained at test time [59]. The one-vs-rest linear SVMs are reported to increase the discriminative power of the feature space.

6.1.5 FaceNet [29]

Instead of training a face recognition system in the form of a conventional classifier, FaceNet implements a system which directly maps the input face thumbnails to the compact Euclidean space. The Euclidean space is generated such that the l2 distance between all faces of the same identity is small, whereas the l2 distance between a pair of face images from different identities is large. This is enabled by triplet loss which, by definition, aims at minimizing the distance between pairs of same identity while maximizing the distance between pairs of different identities.

The authors used two DNNs, (1) Zeiler Fergus [125] and the (2) GoogleNet [36] architecture. The nets were trained on an in-house dataset of 100–200 million face images of about 8 million different identities. Out of the two nets used, Zeiler Fergus, achieved an impressive LFW accuracy of 99.63% and a 95.12% YTF accuracy.

6.1.6 Baidu [57]

The authors present a network comprising 9 convolutions trained with triplet loss. The system reports a near-perfect LFW verification accuracy. The authors conclude that triplet loss, compared to multi-class classification, is more suitable for face verification and retrieval problems.

6.1.7 DLIB [64]

Dlib [64] is library written in C++ which provides software components targeting specialities like data mining, machine learning, image processing, and linear algebra. The library includes a face recognition component that uses a modified version of ResNet-34 [70] to obtain a unique embedding for each face thumbnail. The output feature vectors are of 128 numerical dimensions and the network is trained using triplet loss. The network has been trained on a dataset of 3 million images.

The face recognition component of the Dlib library employs transfer learning to offer flexibility to the user to provide an annotated dataset against which the probe face image/video is compared to. During the enrolment process, the pre-trained model generates vectors for the annotated face images and are stored. During the recognition process, the Euclidean distance between the probe feature vector and each of the stored gallery feature vectors is calculated. During the classification, if the calculated distance lies below a pre-defined threshold, the two faces are considered to be of the same identity. This implementation identifies one or more subjects as possible identity of the unknown face.

6.1.8 OpenFace [65]

OpenFace [65] is a face recognition system open sourced under the Apache 2.0 license. The system was developed with the purpose of bridging the gap between the publicly available face recognition systems and the state-of-the-art high performing private systems. The system is based on concepts introduced in GoogleNet [36] and FaceNet [29].

OpenFace uses a modified version of nn4 network from GoogleNet which was also used in FaceNet. The DNN is trained using triplet loss. The output feature vectors obtained from this trained model are of 128 numerical dimensions.

The face classification is carried out by subject specific modelling approach using a linear SVM. Given the labeled face images of train data, the system generates feature vectors for each face. Then, the feature vectors are fed to the SVM which creates a model based on face feature vectors. When provided with a facial feature vector of an unknown face image, the SVM model classifies the unknown face.

Table 4 Important milestones in face recognition with corresponding LFW verification accuracies

Year	Publication	DCNN architecture	Loss function	Train data	LFW (%)
2014	DeepFace [27]	9 layer deep CNN	Softmax	Private dataset	97.35
2014	DeepId [54]	9 layer deep CNN	Softmax	CelebFaces+	97.45
2014	DeepID2 [55]	9 layer deep CNN	Softmax, contrastive	CelebFaces [106]	99.15
2014	DeepID2+ [56]	9 layer deep CNN	Softmax , contrastive	CelebFaces+,	99.47
				WDRef [129]	99.53
2015	VGGFace [37]	VGGNet [35]	Softmax , triplet	VGGFace [37]	98.95
2015	DeepID3 [28]	GoogleNet [36],	Softmax , contrastive	CelebFaces+ ,	99.53
		VGGNet [35]		WDRef [129]	99.53
2015	FaceNet [29]	GoogleNet [36],	Triplet	Private dataset [29]	99.63
		Zeiler_Fergus [125]			
2015	Baidu [57]	DCNN with 9 convolutions	Triplet	Private dataset	99.77
2018	SphereFace [60]	ResNet-64 [70]	SphereFace [60]	CASIA-WebFace [88]	99.42
2018	CosFace [61]	ResNet-64 [70]	CosFace [61]	CASIA-WebFace [88]	99.33
				Private dataset	99.73
2019	ArcFace [38]	ResNet-100 [70]	ArcFace [38]	ms1m	99.83
Open Source Implementations					
-	DLIB library	ResNet-34 [70]	Triplet	VGGFace [37],	99.38
				FaceScrub [108]	
2016	OpenFace [65]	GoogleNet [36]	Triplet	CASIA-WebFace [88],	92.92
				FaceScrub [108]	
-	FaceNet_Re [66]	Inception-ResNet-v1 [71]	Softmax	VGGFace2 [78]	99.65
		Inception-ResNet-v1 [71]	Softmax	CASIA-WebFace [88]	99.05
		Inception-ResNet-v1 [71]	Centre [79]	VGGFace2 [78]	99.2

6.1.9 FaceNet : re-implementation (FaceNet_Re) [66]

This openly accessible face recognition system is a modified re-implementation of FaceNet [29]. The system provides three pre-trained models of Inception ResNet V1 architecture, trained with varying loss functions and train datasets. As seen in Table 4, the model trained with softmax loss and VGGFace2 reported the highest LFW accuracy out of the three.

Similar to DeepId series, once trained, the inference network which is the network omitting the top layer is used as the pre-trained model generate feature vectors of 512 numerical dimensions. Similar to OpenFace implementation, an SVM classifier is used for classification task.

6.1.10 SphereFace [60] and CosFace [61]

SphereFace and CosFace are two face recognition systems which were used to introduce SphereFace loss and CosFace loss respectively. Both systems use the ResNet-64 architecture and is trained on CASIA-WebFace. Additionally, CosFace trains the system with another private dataset and reports a higher performance.

6.1.11 ArcFace [38]

ArcFace, which is a quite recent publication, implements a series of DNNs (ResNet-100, ResNet-50 and ResNet-34) along with the ArcFace loss. This system outputs a 512-dimensional feature vector for face images. The DNNs were trained on a modified version of Ms Celeb dataset (ms1m). In a series of experimental results, the authors show that this implementation outperforms majority of the reported state-of-the-art results.

6.1.12 Neural aggregation network (NAN) [58]

NAN is a system designed for video face recognition. It comprised a deep network and an aggregation module. The deep network generates feature vectors for faces in video frames. The aggregation module aggregates the feature vectors to form a single feature inside the convex hull spanned by them. This aggregation is invariant to the image order and hence does not utilize the temporal information across video frames. The network used in the paper is of GoogLeNet [36] architecture with the batch normalization [73]. Face verification is carried out with a Siamese NAN structure with two NANs trained with contrastive loss. Face identification is carried out by adding a fully connected layer on top of the NAN for softmax loss. The train dataset uses about 3M face images of 50K identities from the Internet.

6.1.13 Bilinear CNNs (B-CNN) [62]

The system uses a symmetric bilinear-CNN model, comprising two Imagenet-pretrained "M-net" models from VGG's MatConvNet [126]. The models are fine-tuned with FaceScrub dataset. One-versus-rest linear SVM classifiers are trained on the gallery set during experiments.

6.1.14 DCNNmanual+metric [63]

The paper presents an end-to-end system for face verification. The authors train a DCNN with 10 convolutional layers, 5 pooling layers, and 1 fully connected layer with CASIA-WebFace dataset [88]. The system uses joint Bayesian metric learning [127, 128] for face verification. Out of presented deep nets, the network named DCNNmanual+metric yields the best performance. DCNNmanual+metric uses the model trained on CASIA-WebFace

dataset further fine-tuned using the IJB-A [39] and its extended version Janus Challenging set 2 dataset. The system uses cosine distance as a measure of similarity between faces. Manual stands for using training data with manual annotation and metric stands for applying metric learning to compute similarity.

6.2 Study 2: Comparative performance analysis

6.2.1 LFW (2007)

LFW has been the commodity benchmark for face verification, over the last decade. Table 4 presents the summary of recent milestones in face recognition alongside the reported LFW accuracy.

The reported high accuracies on LFW indicates that the benchmark has reached saturation, creating requirement for advanced benchmarks. This near-perfect performance at LFW has been explained by Klare et al. [39], in terms of the nature of the face detector used. This commodity face detector, despite having attractive features like being scalable and real-time efficiency, is not resilient to variations in visual data. Once the faces are mined using this detector, variations like pose, occlusion, and illumination are filtered. The clear and good quality images of frontal pose makes it more convenient to the face recognition systems, thus overlooking the probable challenges in advanced applications like intelligent surveillance. In comparison to the face recognition results reported on larger benchmarks like MegaFace (Table 5), dataset size can be identified as a second factor that enables higher accuracy on LFW.

6.2.2 MegaFace

MegaFace challenge advocates evaluation of deep face recognition at the hand of million distractors. The aim of the benchmark is to scale with the real-world applications that usually involve recognizing a face at a planetary scale.

"Algorithms that achieve above 95% performance on LFW (equivalent of 10 distractors in our plots), achieve 35–75% identification rates with 1 million distractors," reports MegaFace. Accounting for the reported results on this benchmark, Google's FaceNet which achieved near perfect LFW accuracy has recorded an accuracy level of 70%s on MegaFace. The other noteworthy results were of a commercial system named NTechLab. While the reported situation in 2014–2015 was not perfect nor impressive, the years that followed reported progress in recognition results [60, 61]. The recent results reported by ArcFace [38] indicate an impressive near perfect accuracy on MegaFace

Table 5 Face identification and verification evaluation of different methods on MegaFace Challenge1 using FaceScrub as the probe set

Publication	Id (%)	Ver (%)
FaceNet v8	70.49	-
NTechLAB	73.30	-
SphereFace	72.729	85.561
CosFace (single-patch)	72.729	96.65
CosFace (3-patch ensemble)	74.11	97.96
ArcFace	77.06	96.98
ArcFace R	98.35	98.48

Id rank 1 face identification accuracy with 1M distractors, *Ver* face verification TAR at 10-6 FAR, *R* data refinement on both probe set and 1M distractors

benchmark. Please refer to Table 5 for a summary of identification and verification results on MegaFace.

6.2.3 IJB (2015)

Table 6 presents a summary of reported face recognition results on IJB benchmark. While the reported results on this benchmark are comparatively higher than those on MegaFace, the results are not perfect, nor near-perfect. Hence, these results are an indication that the face recognition is challenged by complications in unconstrained data. A noteworthy fact regarding this benchmark is that, since the dataset includes multiple imagery for a single recognition, ideally, the system should include a mechanism to fully exploit the excess information. While the authors of the dataset suggest subject specific modelling, systems like B-CNN have employed other approaches like result pooling.

6.3 Study 3: Experimental analysis

Bianco et al. [77] presents an experimental analysis of DCNN frameworks for image recognition. Here, experiments on all systems are carried out on the same computational resources. Hence, it provides an unbiased comparison of strengths and weaknesses of the frameworks. Inspired by the work of Bianco et al., this experimental study analyzes the performance of three open-source face recognition systems (DLIB library, OpenFace, and FaceNet_Re) in terms of recognition accuracy and speed. The systems in comparison use three main deep network architectures discussed in this survey; ResNet, GoogleNet, and Inception-ResNet and the two main classification approaches, subject-specific modeling with SVM and template learning.

OpenFace and DLIB uses HoG face detector [116] while FaceNet_Re uses MTCNN face detector [115]. To avoid dependencies from the detectors, only the faces detected by both algorithms were considered in the experiment. Taking into account the dependencies from different classification approaches, the two systems that used subject-specific modeling with linear SVMs (OpenFace and FaceNet_Re) were modified to perform template comparison in a similar manner to that of DLIB (nearest neighbor based on euclidean distance). In addition, the results from the original SVM implementation was also reported for comparison.

Depending on the use case, the recognition could be from Still images to Still images (S2S), from Video to Video (V2V), and from Still images to Video (S2V). Several benchmarks have addressed the first two approaches; LFW and MegaFace address S2S, and IJB addresses the combination of S2V, S2S, and V2V tests. While some benchmarks have made efforts to address S2V, these are problem specific datasets with some form of bias [130]. Hence, the experiment measures S2V recognition with LFW dataset as the set of gallery images and selected videos of YTF dataset as the probe videos. The experiment measures the rank 1 recognition accuracy with increasing gallery sizes. In addition, the average time taken by the system to run the forward pass on a single face thumbnail is compared.

Figure 9 plots the recognition accuracy against the gallery size. The graph depicts two observations: (1) comparing the performance of the three systems with template learning as the classifier, FaceNet_Re TL and OpenFace TL show performance decrease with the gallery size; however, DLIB system shows considerable stability against the growth of gallery; and (2) comparing the performance of the same system with SVM and template

Table 6 Performance evaluation on the IJB-A dataset

Publication	**1:1 verification TAR**		**1:N identification TPIR**			
	FAR = 0.001	**FAR = 0.01**	**FPIR = 0.01**	**FPIR = 0.01**	**Rank 1**	**Rank 5**
VGGFace [37]	-	0.461 ± 0.077	0.670 ± 0.031	0.913 ± 0.011	-	
Template [59]	0.939 ± 0.013	0.979 ± 0.004	0.774 ± 0.049	0.882 ± 0.016	0.928 ± 0.010	0.977 ± 0.004
NAN [58]	0.941 ± 0.008	0.978 ± 0.003	0.817 ± 0.041	0.917 ± 0.009	0.958 ± 0.005	0.980 ± 0.005
B-CNN [62]	-	-	0.143 ± 0.027	0.341 ± 0.032	0.588 ± 0.020	0.796 ± 0.017
DCNNmanual+metric [63]	-	0.787 ± 0.043	-	-	0.852 ± 0.018	0.937 ± 0.010

For verification, the true accept rates (TAR) vs. false-positive rates (FAR) are reported. For identification, the true-positive identification rate (TPIR) vs. false-positive identification rate (TPIR) and the rank-N accuracies are presented

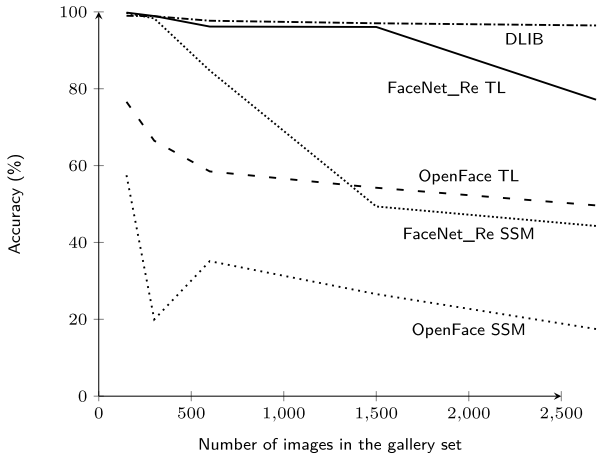

Fig. 9 Face recognition accuracy against the gallery size. SSM, subject-specific modeling; TL, template learning. The gallery contains three images per subject and ranges from 150 images to 2700 images, i.e., 50–900 subjects

learning classifiers, in both the instances (OpenFace and FaceNet_Re), the SVM is effective with limited number of subjects, but the performance drops drastically as the number of subjects increases. And one-to-one template learning is comparatively more stable against larger gallery sets. Since many studies have encouraged the use of subject-specific modeling to better utilize all the available information from multiple visual data [27, 39–41, 59], it is important to properly analyze the strengths and weaknesses of different modeling approaches. The popularity of SVM in image classification can be explained by its ability to scale well with high dimensional data [131–133]. Although this works well when provided with small number of classes, increased number of classes with limited train data per class could complicate the process of finding the separation hyperplane.

Table 7 reports the average time taken by each system to run the DCNN model on a single face thumbnail, as recorded on an Intel Core i7-7740X CPU @ 4.30GHz. The times reflect the underlying computational complexity involved in feature extraction from raw pixels. OpenFace and FaceNet_Re that includes Inception modules in the framework have reported lesser forward pass time in comparison to the DLIB model. Among the limited records in literature on computational efficiency, DeepFace reports an 0.18-s feature extraction time on a single core Intel 2.2GHz CPU and FaceNet reports a 30 ms/image on a mobile phone with a small NN which is reported to have lesser, yet sufficient-for-face-clustering accuracy.

7 Open issues

Starting from face verification with high-quality data, face recognition has advanced over the recent years to address complicated scenarios like face recognition in unconstrained images and video face recognition. Simultaneously, face recognition benchmarks like IJB and MegaFace have aimed to replicate real-world applications. While the reports

Table 7 Feature extraction time per face

System	NN	Forward pass (s)
OpenFace	GoogleNet [36]	0.08
DLIB	ResNet-34 [70]	0.15
FaceNet_Re	Inception-ResNet-v1 [71]	0.01

indicates a continuous progress, there are some un-addressed issues in terms of face recognition systems and benchmarks.

7.1 A comparative analysis for face recognition accuracy

Several studies have carried out experimental evaluations comparing state-of-the-art DCNN frameworks for image classification [68, 76, 134]. These experiments provide unbiased comparisons of the systems. This is particularly important in situations where all the systems are not evaluated on the same benchmark. The condition applies to face recognition as well. While almost all the face recognition systems provide the LFW accuracy, systems that were implemented prior to benchmarks like IJB and MegaFace do not provide evaluation results on them. Hence, there exists the necessity for these systems to be evaluated under a common benchmark.

7.2 A comparative analysis for computational complexity

While the studies report the recorded accuracy, only limited publications report the associated computational complexity. Despite offline processing being generally flexible on computational complexity, it is one of the most critical requirements in real-time applications. Hence, there exists the need for a comparative analysis of the deep face recognition systems with respect to performance indices like computational complexity, memory consumption, and inference time.

7.3 End-to-end systems

Majority of the studies and benchmarks tend to isolate face recognition as an individual discipline and hence do not provide sufficient insights on critical issues arising from inevitable integration with modules like face detection (e.g., false recognitions resulting from false detections). Despite limited studies [63, 135, 136], end-to-end face recognition is still an open research.

7.4 Multi-model face recognition

Most of the deep face recognition systems generate a single feature embedding for each face. This approach consider holistic features and does not contemplate component level features. Several studies have aimed to implement multi-model face recognition systems to gain optimum use of diverse information in a face image [137–139]. Several studies have made efforts to perform fusion of multiple descriptors across face [140–142]. These systems show that despite the possibility of increased computational complexity, the multi-model systems can yield positive results. Hence, this study an be improved targeting applications that require offline processing.

7.5 Multi-face recognition and tracking

The benchmarks and systems for video face recognition portray the problem as face recognition on a set of face images per subject [39–41]. These benchmarks does not evaluate face tracking. Nonetheless, face tracking is of vital importance in multi-face recognition in videos. In this scenario, the pixel level information in a video frame and the temporal information across video frames can be fused for an improved result. While face tracking and face clustering have been studies as a separate discipline [143–146], in practical applications, they are applied along with face recognition. Hence, evaluating

the state-of-the-art face recognition systems along with face tracking can be a possible research with practical use.

7.6 Ensemble of deep learning and traditional face descriptors

While deep face descriptors are becoming the main feature representations for face recognition, traditional visual appearance descriptors can be used as an additional informational guidance. Recent developments have demonstrated effective usage of traditional visual descriptors in image processing tasks such as image semantic learning [147] and text mining in complex background images [148]. Exploring their effectiveness as an ensemble of deep learning could be possible future research.

7.7 Video face recognition

Frame-wise face recognition, feature aggregation across frames, and score pooling across frames are popular approaches of video face recognition. The first approach provides a crisp classification output that the probe face belongs to identity x. Unconstrained videos where faces are subjected to motion blur and other factors like partial faces due to pose might require aggregation of several partial truths into a higher truth. Despite score pooling and feature aggregation being straightforward aggregations across video frames, there exists room for sophisticated algorithms like inference based on fuzzy logic. They can be adapted from research work of similar disciplines like image annotation [149]. Through this mechanism, a degree of certainty can be calculated to the classification output, against factors like the quality of the image and the fraction of face visible.

While temporal attention has proven to be effective in video face recognition, research disciplines like video captioning has shown improved performance by including spatial affinities in the attention [150]. Hence, spatial-temporal attention emerges as a possible research for video face recognition.

7.8 Application-specific designs

The expected functionality of face recognition varies with application. A face recognition application designed for intelligent surveillance where the cost of false alarms (registered individuals recognized as possible intruders) is high and the cost of missed alarms (possible intruders recognized as registered) is even higher should strive for minimum FPIR with reasonable flexibility on false-negative rejections. In contrast, applications that detects persons of interest is expected to have minimum FNIR (person of interest recognized as unknown) with reasonable flexibility of FPIR (false alarm where a regular individual is identified as a person of interest). Hence, the need for scenario specific designs and benchmarks cannot be overlooked.

7.9 Basic challenges for face recognition

Despite many architectural enhancements and diversified datasets, face recognition still has scope for improvement in terms of elementary complications arising from visual variations like pose, expression, occlusion, and illumination. In addition to direct studies like expression invariant face recognition [151], face recognition under occlusion [152], or illumination face recognition [52], tools like video segmentation [153, 154] and region of interest extraction [155] can provide potential indirect assistance for face recognition on noisy imagery.

Regardless of the varying modeling approaches and application specific fine-tuning, face recognition has mainly been influenced by DCNN frameworks, loss functions, classification algorithms, and train data. The continuous advancement of deep network architectures in image classification generates networks adaptable for face recognition. The study of Dong et al. [156], Bruna and Mallat [157], and Hankins et al. [158] are some image classification networks with prospect for face classification. Hence, face recognition will remain an active research striving for sophisticated frameworks.

8 Conclusion

This survey has presented the origin and evolution and a comparative analysis of 18 face recognition systems. Through this, the survey aims provide an informational guidance to simulate future research. In doing so, the paper has analyzed the performance of the systems in terms of benchmark results reported on three benchmarks which addresses different aspects of face recognition and an experimental study. Additionally, the survey has discussed the open issues in face recognition along with a note on possible future research.

Acknowledgements
The research activity leading to the publication has been partially funded by the European Union Horizon 2020 research and innovation program under grant agreement No. 787123 (PERSONA RIA project).

Authors' contributions
SWA drafted the manuscript, acquired and analyzed the data, and carried out the experimental study presented in the manuscript. EI made substantial contributions to the conception and design for the research and experimental study. EI revised the manuscript completely and critically for important intellectual content. Each author has given final approval of the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The authors read and approved the final manuscript.

Funding
The research activity leading to the publication has been partially funded by the European Union Horizon 2020 research and innovation program under grant agreement No. 787123 (PERSONA RIA project).

Availability of data and materials
The LFW dataset analysed during the current study are available in the Labeled Faces in the Wild repository, http://vis-www.cs.umass.edu/lfw/ [31]. The YTF dataset analysed during the current study are available in the YouTube Faces DB repository, https://www.cs.tau.ac.il/~wolf/ytfaces/ [96]. The OpenFace system analysed during the current study are available in the OpenFace repository, https://cmusatyalab.github.io/openface/ [65]. The archived version referenced in the manuscript is corresponding to the release 0.2.1. The project is licensed under the Apache 2.0 License. The FaceNet_Re system analysed during the current study are available in the FaceNet repository, https://github.com/davidsandberg/facenet [66]. The archived version referenced in the manuscript is https://github.com/davidsandberg/facenet/tree/096ed770f163957c1e56efa7feeb194773920f6e. 096ed770f163957c1e56efa7feeb194773920f6e. The project is free and open source licensed under the Apache MIT License. The DLIB system analyzed during the current study are available in the FaceNet repository, https://github.com/ageitgey/face_recognition [64]. The archived version referenced in the manuscript is corresponding to version 1.2.3. The project is free and open source licensed under the Apache MIT License.

Competing interests
The authors declare that they have no competing interests.

Received: 27 October 2019 Accepted: 4 May 2020
Published online: 29 June 2020

References

1. A. M. Burrows, J. F. Cohn, in *Encyclopedia of Biometrics, Second Edition*, Comparative anatomy of the face, (2015), pp. 313–321. https://doi.org/10.1007/978-1-4899-7488-4_190
2. R. Chellappa, C. L. Wilson, S. Sirohey, Human and machine recognition of faces: a survey. Proc. IEEE. **83**(5), 705–741 (1995). https://doi.org/10.1109/5.381842
3. Y. Wang, T. Bao, D. Ding, M. Zhu, in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Face recognition in real-world surveillance videos with deep learning method, (2017), pp. 239–243. https://doi.org/10.1109/ICIVC.2017.7984553

4. S. Degadwala, S. Pandya, V. Patel, S. Shah, U. Doshi, in *International Conference on Recent Trends in Engineering, Science Technology - (ICRTEST 2016)*, A review on real time face tracking and identification for surveillance system, (2016), pp. 1–5. https://doi.org/10.1049/cp.2016.1477
5. , in *International Civil Aviation Organization (ICAO) Doc 9303 vol. 2*, Machine readable travel documents, (2006)
6. R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, F. Scotti, G. Sforza, Biometric recognition in automated border control: a survey. ACM Comput. Surv. **49**(2), 24–12439 (2016). https://doi.org/10.1145/2933241
7. J. Y. Choi, W. De Neve, K. N. Plataniotis, Y. M. Ro, Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks. IEEE Trans. Multimed. **13**(1), 14–28 (2011). https://doi.org/10.1109/TMM.2010.2087320
8. Q. Xu, M. Mukawa, L. Li, J. H. Lim, C. Tan, S. C. Chia, T. Gan, B. Mandal, in *Proceedings of the 6th Augmented Human International Conference, AH '15*, Exploring users' attitudes towards social interaction assistance on google glass (ACM, New York, NY, USA, 2015), pp. 9–12. https://doi.org/10.1145/2735711.2735831. http://doi.acm.org/10.1145/2735711.2735831
9. B. Mandal, R. Y. Lim, P. Dai, M. R. Sayed, L. Li, J. H. Lim, *Trends in machine and human face recognition*. (M. Kawulok, M. E. Celebi, B. Smolka, eds.) (Springer, Cham, 2016), pp. 145–187. https://doi.org/10.1007/978-3-319-25958-1_7
10. B. Mandal, in *2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Face recognition: perspectives from the real world, (2016), pp. 1–5. https://doi.org/10.1109/ICARCV.2016.7838675
11. X. Zhang, Y. Gao, Face recognition across pose: a review. Pattern Recogn. **42**(11), 2876–2896 (2009). https://doi.org/10.1016/j.patcog.2009.04.017
12. K. Jia, S. Gong, *Face sample quality*. (S. Z. Li, A. K. Jain, eds.) (Springer, Boston, MA, 2015), pp. 522–526. https://doi.org/10.1007/978-1-4899-7488-4_86
13. C. Conde, I. M. de Diego, E. Cabello, in *E-business and telecommunications*. ed. by M. S. Obaidat, J. L. Sevillano, and J. Filipe, Face recognition in uncontrolled environments, experiments in an airport (Springer, Berlin, Heidelberg, 2012), pp. 20–32
14. W. W. Bledsoe, The model method in facial recognition. Technical Report, PRI 15, Panoramic Research (1964). Inc., Palo Alto, CA, California
15. W. Bledsoe, Man-machine facial recognition: Report on a large-scale experiment. Panoramic Research (1966). Inc., Palo Alto, CA
16. A. Serrano, I. M. de Diego, C. Conde, E. Cabello, Recent advances in face biometrics with gabor wavelets: a review. Pattern Recogn. Lett. **31**(5), 372–381 (2010). https://doi.org/10.1016/j.patrec.2009.11.002
17. P. S. Penev, J. J. Atick, Local feature analysis: a general statistical theory for object representation, (1996). https://doi.org/10.1088/0954-898x_7_3_002
18. T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001). https://doi.org/10.1109/34.927467
19. L. Wiskott, J.-. Fellous, N. Kruger, C. von der Malsburg, in *Proceedings of International Conference on Image Processing, vol. 1*, Face recognition by elastic bunch graph matching, (1997), pp. 129–1321. https://doi.org/10.1109/ICIP.1997.647401
20. T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006). https://doi.org/10.1109/TPAMI.2006.244
21. S. Chen, S. Mau, M. T. Harandi, C. Sanderson, A. Bigdeli, B. C. Lovell, Face recognition from still images to video sequences: a local-feature-based framework. EURASIP J. Video Process. **2011**(1), 790598 (2010). https://doi.org/10.1155/2011/790598
22. A. Rattani, D. R. Kisku, M. Bicego, M. Tistarelli, in *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, Feature level fusion of face and fingerprint biometrics, (2007), pp. 1–6. https://doi.org/10.1109/BTAS.2007.4401919
23. N. Dalal, B. Triggs, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1*, Histograms of oriented gradients for human detection, (2005), pp. 886–8931. https://doi.org/10.1109/CVPR.2005.177
24. M. Turk, A. Pentland, Eigenfaces for recognition. J. Cogn. Neurosci. **3**(1), 71–86 (1991). https://doi.org/10.1162/jocn.1991.3.1.71. PMID: 23964806. https://doi.org/10.1162/jocn.1991.3.1.71
25. P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997). https://doi.org/10.1109/34.598228
26. J. Galbally, C. McCool, J. Fierrez, S. Marcel, J. Ortega-Garcia, On the vulnerability of face verification systems to hill-climbing attacks. Pattern Recogn. **43**(3), 1027–1038 (2010). https://doi.org/10.1016/j.patcog.2009.08.022
27. Y. Taigman, M. Yang, M. Ranzato, L. Wolf, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Deepface: closing the gap to human-level performance in face verification, (2014), pp. 1701–1708. https://doi.org/10.1109/CVPR.2014.220
28. Y. Sun, D. Liang, X. Wang, X. Tang, Deepid3: face recognition with very deep neural networks. CoRR. **abs/1502.00873** (2015). 1502.00873
29. F. Schroff, D. Kalenichenko, J. Philbin, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Facenet: a unified embedding for face recognition and clustering, (2015), pp. 815–823. https://doi.org/10.1109/CVPR.2015.7298682
30. R. Jafri, H. Arabnia, A survey of face recognition techniques. J. Inf. Process. Syst. (JIPS). **5**, 41–68 (2009). https://doi.org/10.3745/JIPS.2009.5.2.041
31. G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, in *Workshop on faces in 'real-life' images: detection, alignment, and recognition*, Labeled faces in the wild: a database for studying face recognition in Unconstrained Environments (Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Marseille, France, 2008). https://hal.inria.fr/inria-00321923
32. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proc. IEEE. **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791

33. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
34. A. Krizhevsky, I. Sutskever, G. E. Hinton, in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12*, Imagenet classification with deep convolutional neural networks (Curran Associates Inc., USA, 2012), pp. 1097–1105. http://dl.acm.org/citation.cfm?id=2999134.2999257
35. K. Simonyan, A. Zisserman, in *International Conference on Learning Representations*, Very deep convolutional networks for large-scale image recognition, (2015)
36. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Going deeper with convolutions, (2015), pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594
37. O. M. Parkhi, A. Vedaldi, A. Zisserman, in *Proceedings of the British Machine Vision Conference (BMVC)*. ed. by M. W. J. Xianghua Xie, G. K. L. Tam, Deep face recognition (BMVA Press, 2015), pp. 41–14112. https://doi.org/10.5244/C.29.41
38. J. Deng, J. Guo, S. Zafeiriou, Arcface: additive angular margin loss for deep face recognition. CoRR. **abs/1801.07698** (2018). 1801.07698
39. B. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. E. Allen, P. Grother, A. Mah, M. Burge, A. K. Jain, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a, (2015), pp. 1931–1939. https://doi.org/10.1109/cvpr.2015.7298803
40. C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, P. Grother, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Iarpa janus benchmark-b face dataset, (2017), pp. 592–600. https://doi.org/10.1109/CVPRW.2017.87
41. B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, P. Grother, in *2018 International Conference on Biometrics (ICB)*, Iarpa janus benchmark - c: face dataset and protocol, (2018), pp. 158–165. https://doi.org/10.1109/ICB2018.2018.00033
42. I. Masi, Y. Wu, T. Hassner, P. Natarajan, in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Deep face recognition: a survey, (2018), pp. 471–478. https://doi.org/10.1109/sibgrapi.2018.00067
43. B. Mandal, Z. Wang, L. Li, A. A. Kassim, Performance evaluation of local descriptors and distance measures on benchmarks and first-person-view videos for face identification. Neurocomputing. **184**, 107–116 (2016). https://doi.org/10.1016/j.neucom.2015.07.121. RoLoD: Robust Local Descriptors for Computer Vision 2014
44. K. W. Bowyer, K. I. Chang, P. J. Flynn, A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. Comput. Vis. Image Underst. **101**, 1–15 (2006)
45. A. Scheenstra, A. Ruifrok, R. C. Veltkamp, in *Audio- and video-based biometric person authentication*. ed. by T. Kanade, A. Jain, and N. K. Ratha, A survey of 3D face recognition methods (Springer, Berlin, Heidelberg, 2005), pp. 891–899
46. M. Wang, W. Deng, Deep face recognition: a survey. CoRR. **abs/1804.06655** (2018). 1804.06655
47. S. Zhou, S. Xiao, 3D face recognition: a survey. Hum. Centric Comput. Inf. Sci. **8**(1), 35 (2018). https://doi.org/10.1186/s13673-018-0157-2
48. C. T. Ferraz, J. H. Saito, in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, WebMedia '18*, A comprehensive analysis of local binary convolutional neural network for fast face recognition in surveillance video (ACM, New York, NY, USA, 2018), pp. 265–268. https://doi.org/10.1145/3243082.3267444
49. T. Patel, B. Shah, in *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, A survey on facial feature extraction techniques for automatic face annotation, (2017), pp. 224–228. https://doi.org/10.1109/ICIMIA.2017.7975607
50. B. Prihasto, S. Choirunnisa, M. I. Nurdiansyah, S. Mathulaprangsan, V. C. Chu, S. Chen, J. Wang, in *2016 International Conference on Orange Technologies (ICOT)*, A survey of deep face recognition in the wild, (2016), pp. 76–79. https://doi.org/10.1109/ICOT.2016.8278983
51. C. Ding, D. Tao, A comprehensive survey on pose-invariant face recognition. CoRR. **abs/1502.04383** (2015). 1502.04383
52. M. A. Ochoa-Villegas, J. A. Nolazco-Flores, O. Barron-Cano, I. A. Kakadiaris, Addressing the illumination challenge in two-dimensional face recognition: a survey. IET Comput. Vis. **9**(6), 978–992 (2015). https://doi.org/10.1049/iet-cvi.2014.0086
53. R. Tyagi, G. Tomar, N. Baik, A survey of unconstrained face recognition algorithm and its applications. Int. J. Secur. Appl. **10**, 369–376 (2016). https://doi.org/10.14257/ijsia.2016.10.12.30
54. Y. Sun, X. Wang, X. Tang, in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, Deep learning face representation from predicting 10,000 classes (IEEE Computer Society, Washington, DC, USA, 2014), pp. 1891–1898. https://doi.org/10.1109/CVPR.2014.244
55. Y. Sun, Y. Chen, X. Wang, X. Tang, in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14*, Deep learning face representation by joint identification-verification (MIT Press, Cambridge, MA, USA, 2014), pp. 1988–1996. http://dl.acm.org/citation.cfm?id=2969033.2969049
56. Y. Sun, X. Wang, X. Tang, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Deeply learned face representations are sparse, selective, and robust, (2015), pp. 2892–2900. https://doi.org/10.1109/cvpr.2015.7298907. https://app.dimensions.ai/details/publication/pub.1095455292andhttp://arxiv.org/pdf/1412.1265
57. J. Liu, Y. Deng, T. Bai, C. Huang, Targeting ultimate accuracy: face recognition via deep embedding. CoRR. **abs/1506.07310** (2015). 1506.07310
58. J. Yang, P. Ren, D. Chen, F. Wen, H. Li, G. Hua, *Neural aggregation network for video face recognition*, (2016), pp. 5216–5225. https://doi.org/10.1109/cvpr.2017.554
59. N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, A. Zisserman, in *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, Template adaptation for face verification and identification, (2017), pp. 1–8. https://doi.org/10.1109/FG.2017.11

60. W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Sphereface: deep hypersphere embedding for face recognition, (2017), pp. 6738–6746. https://doi.org/10.1109/CVPR.2017.713
61. H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Cosface: large margin cosine loss for deep face recognition, (2018), pp. 5265–5274. https://doi.org/10.1109/CVPR.2018.00552
62. A. Roy Chowdhury, T. Lin, S. Maji, E. G. Learned-Miller, Face identification with bilinear CNNs. CoRR. **abs/1506.01342** (2015). 1506.01342
63. J. Chen, R. Ranjan, A. Kumar, C. Chen, V. M. Patel, R. Chellappa, in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, An end-to-end system for unconstrained face verification with deep convolutional neural networks, (2015), pp. 360–368. https://doi.org/10.1109/ICCVW.2015.55
64. High quality face recognition with deep metric learning. http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html. Accessed 20 May 2019
65. B. Amos, B. Ludwiczuk, M. Satyanarayanan, Openface: a general-purpose face recognition library with mobile applications, CMU-CS-16-118, CMU School of Computer Science, (2016)
66. Face recognition using TensorFlow. https://github.com/davidsandberg/facenet. Accessed 20 May 2019
67. I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, E. Brossard, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, The megaface benchmark: 1 million faces for recognition at scale, (2016), pp. 4873–4882. https://doi.org/10.1109/CVPR.2016.527
68. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature. **521**, 436–44 (2015). https://doi.org/10.1038/nature14539
69. M. Lin, Q. Chen, S. Yan, Network in network. CoRR. **abs/1312.4400** (2013)
70. K. He, X. Zhang, S. Ren, J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Deep residual learning for image recognition, (2016), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
71. C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17*, Inception-v4, inception-resnet and the impact of residual connections on learning (AAAI Press, 2017), pp. 4278–4284. http://dl.acm.org/citation.cfm?id=3298023.3298188
72. F. Song, J. Dongarra, in *Proceedings of the 28th ACM International Conference on Supercomputing, ICS '14*, Scaling up matrix computations on shared-memory manycore systems with 1000 CPU cores (ACM, New York, NY, USA, 2014), pp. 333–342. https://doi.org/10.1145/2597652.2597670
73. S. Ioffe, C. Szegedy, in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15*, Batch normalization: accelerating deep network training by reducing internal covariate shift (JMLR.org, 2015), pp. 448–456. http://dl.acm.org/citation.cfm?id=3045118.3045167
74. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, in *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Rethinking the inception architecture for computer vision, (2016). https://doi.org/10.1109/CVPR.2016.308
75. T. Lin, A. RoyChowdhury, S. Maji, in *2015 IEEE International Conference on Computer Vision (ICCV)*, Bilinear CNN models for fine-grained visual recognition, (2015), pp. 1449–1457. https://doi.org/10.1109/ICCV.2015.170
76. A. Canziani, A. Paszke, E. Culurciello, An analysis of deep neural network models for practical applications. CoRR. **abs/1605.07678** (2016). 1605.07678
77. S. Bianco, R. Cadène, L. Celona, P. Napoletano, Benchmark analysis of representative deep neural network architectures. CoRR. **abs/1810.00736** (2018). 1810.00736
78. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, in *International Conference on Automatic Face and Gesture Recognition*, Vggface2: a dataset for recognising faces across pose and age, (2018). https://doi.org/10.1109/fg.2018.00020
79. Y. Wen, K. Zhang, Z. Li, Y. Qiao, in *Computer vision – ECCV 2016*. ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling, A discriminative feature learning approach for deep face recognition (Springer, Cham, 2016), pp. 499–515
80. J. Deng, Y. Zhou, S. Zafeiriou, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Marginal loss for deep face recognition, (2017), pp. 2006–2014. https://doi.org/10.1109/CVPRW.2017.251
81. X. Zhang, Z. Fang, Y. Wen, Z. Li, Y. Qiao, in *2017 IEEE International Conference on Computer Vision (ICCV)*, Range loss for deep face recognition with long-tailed training data, (2017), pp. 5419–5428. https://doi.org/10.1109/ICCV.2017.578
82. W. Liu, Y. Wen, Z. Yu, M. Yang, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, Large-margin softmax loss for convolutional neural networks (JMLR.org, 2016), pp. 507–516. http://dl.acm.org/citation.cfm?id=3045390.3045445
83. F. Wang, J. Cheng, W. Liu, H. Liu, Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**(7), 926–930 (2018). https://doi.org/10.1109/LSP.2018.2822810
84. B. Chen, W. Deng, J. Du, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Noisy softmax: improving the generalization ability of dcnn via postponing the early softmax saturation, (2017). https://doi.org/10.1109/CVPR.2017.428
85. W. Wan, Y. Zhong, T. Li, J. Chen, Rethinking feature distribution for loss functions in image classification. CoRR. **abs/1803.02988** (2018). 1803.02988
86. X. Qi, L. Zhang, Face recognition via centralized coordinate learning. CoRR. **abs/1801.05678** (2018). 1801.05678
87. Y. Sun, X. Wang, X. Tang, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Sparsifying neural network connections for face recognition, (2016), pp. 4856–4864. https://doi.org/10.1109/CVPR.2016.525
88. D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch. CoRR. **abs/1411.7923** (2014). 1411.7923
89. J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, in *Proceedings of the 6th International Conference on Neural Information Processing Systems, NIPS'93*, Signature verification using a "siamese" time delay neural network (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993), pp. 737–744. http://dl.acm.org/citation.cfm?id=2987189.2987282

90. R. Hadsell, S. Chopra, Y. LeCun, in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2*, Dimensionality reduction by learning an invariant mapping, (2006), pp. 1735–1742. https://doi.org/10.1109/CVPR.2006.100
91. A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CNN features off-the-shelf: an astounding baseline for recognition, (2014), pp. 512–519. https://doi.org/10.1109/CVPRW.2014.131
92. S. J. Pan, Q. Yang, A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
93. K. Kim, Z. Yang, I. Masi, R. Nevatia, G. Medioni, in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Face and body association for video-based face recognition, (2018), pp. 39–48. https://doi.org/10.1109/WACV.2018.00011
94. H. Li, G. Hua, X. Shen, Z. Lin, J. Brandt, in *Computer vision – ACCV 2014*. ed. by D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eigen-pep for video face recognition (Springer, Cham, 2015), pp. 17–33
95. Z. Liu, H. Hu, J. Bai, S. Li, S. Lian, in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Feature aggregation network for video face recognition, (2019). https://doi.org/10.1109/iccvw.2019.00128
96. L. Wolf, T. Hassner, I. Maoz, in *2011 IEEE Conference on Computer Vision and Pattern Recognition*, Face recognition in unconstrained videos with matched background similarity, (2011), pp. 529–534. https://doi.org/10.1109/CVPR.2011.5995566
97. B. Chen, C. Chen, W. H. Hsu, Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Trans. Multimed. **17**(6), 804–815 (2015). https://doi.org/10.1109/TMM.2015.2420374
98. B.-C. Chen, C.-S. Chen, W. H. Hsu, in *Computer vision – ECCV 2014*. ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Cross-age reference coding for age-invariant face recognition and retrieval (Springer, Cham, 2014), pp. 768–783
99. T. Zheng, W. Deng, J. Hu, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Age estimation guided convolutional neural network for age-invariant face recognition, (2017), pp. 503–511. https://doi.org/10.1109/cvprw.2017.77
100. K. Ricanek, T. Tesafaye, in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, Morph: a longitudinal image database of normal adult age-progression, (2006), pp. 341–345. https://doi.org/10.1109/FGR.2006.78
101. T. Zheng, W. Deng, *Cross-pose LFW: a database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01*. (Beijing University of Posts and Telecommunications, 2018)
102. V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, R. Chellappa, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Disguised faces in the wild, (2018), pp. 1–18. https://doi.org/10.1109/CVPRW.2018.00008
103. Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Web-scale training for face identification. CoRR. **abs/1406.5266** (2014). 1406.5266
104. P. J. Phillips, P. Grother, R. Micheals, *Evaluation methods in face recognition*. (S. Z. Li, A. K. Jain, eds.) (Springer, London, 2011), pp. 551–574. https://doi.org/10.1007/978-0-85729-932-1_21
105. P. Viola, M. Jones, in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1*, Rapid object detection using a boosted cascade of simple features, (2001). https://doi.org/10.1109/CVPR.2001.990517
106. Y. Sun, X. Wang, X. Tang, Hybrid deep learning for face verification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 1997–2009 (2016). https://doi.org/10.1109/TPAMI.2015.2505293
107. Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, in *European Conference on Computer Vision, vol. 9907*, Ms-celeb-1m: a dataset and benchmark for large-scale face recognition, (2016), pp. 87–102. https://doi.org/10.1007/978-3-319-46487-9_6
108. H. Ng, S. Winkler, in *2014 IEEE International Conference on Image Processing (ICIP)*, A data-driven approach to cleaning large face datasets, (2014), pp. 343–347. https://doi.org/10.1109/ICIP.2014.7025068
109. G. Panis, A. Lanitis, An Overview of Research Activities in Facial Age Estimation Using the FG-NET Aging Database. **8926**, 737–750 (2015). https://doi.org/10.1007/978-3-319-16181-5_56
110. I. Kemelmacher-Shlizerman, S. Suwajanakorn, S. M. Seitz, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Illumination-aware age progression, (2014), pp. 3334–3341. https://doi.org/10.1109/CVPR.2014.426
111. B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L. Li, The new data and new challenges in multimedia research. CoRR. **abs/1503.01817** (2015). 1503.01817
112. F. H. d.B. Zavan, N. Gasparin, J. C. Batista, L. P. e.Silva, V. Albiero, O. R. P. Bellon, L. Silva, in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, Face analysis in the wild, (2017), pp. 9–16. https://doi.org/10.1109/SIBGRAPI-T.2017.11
113. Y. Wu, Q. Ji, Facial landmark detection: a literature survey. Int. J. Comput. Vis. **127**(2), 115–142 (2019). https://doi.org/10.1007/s11263-018-1097-z
114. Y. Wu, T. Hassner, K. Kim, G. Medioni, P. Natarajan, Facial landmark detection with tweaked convolutional neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 3067–3074 (2018)
115. K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016). https://doi.org/10.1109/LSP.2016.2603342
116. N. Dalal, B. Triggs, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1*, Histograms of oriented gradients for human detection, (2005), pp. 886–8931. https://doi.org/10.1109/CVPR.2005.177
117. O. Çeliktutan, S. Ulukaya, B. Sankur, A comparative study of face landmarking techniques. EURASIP J. Image Video Process. **2013**(1), 13 (2013). https://doi.org/10.1186/1687-5281-2013-13
118. B. Johnston, P. d. Chazal, A review of image-based automatic facial landmark identification techniques. EURASIP J. Image Video Process. **2018**(1), 86 (2018). https://doi.org/10.1186/s13640-018-0324-4
119. X. Lin, Y. Liang, J. Wan, C. Lin, S. Z. Li, Trans. Multimed. IEEE, Region-based context enhanced network for robust multiple face alignment, 1–1 (2019). https://doi.org/10.1109/TMM.2019.2916455

120. R. Weng, J. Lu, Y. Tan, J. Zhou, Learning cascaded deep auto-encoder networks for face alignment. IEEE Trans. Multimed. **18**(10), 2066–2078 (2016). https://doi.org/10.1109/TMM.2016.2591508
121. W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, R. Nevatia, G. Medioni, in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Face recognition using deep multi-pose representations, (2016), pp. 1–9. https://doi.org/10.1109/WACV.2016.7477555
122. A. Bulat, G. Tzimiropoulos, How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). CoRR. **abs/1703.07332** (2017). 1703.07332
123. F. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, G. G. Medioni, Faceposenet: making a case for landmark-free face alignment. CoRR. **abs/1708.07517** (2017). 1708.07517
124. X. Zhu, Z. Lei, X. Liu, H. Shi, S. Z. Li, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Face alignment across large poses: a 3D solution, (2016), pp. 146–155. https://doi.org/10.1109/cvpr.2016.23
125. L. Wolf, T. Hassner, I. Maoz, in *2011 IEEE Conference on Computer Vision and Pattern Recognition*, Face recognition in unconstrained videos with matched background similarity, (2011), pp. 529–534. https://doi.org/10.1109/CVPR.2011.5995566
126. A. Vedaldi, K. Lenc, in *Proceeding of the ACM Int. Conf. on Multimedia*, Matconvnet – convolutional neural networks for matlab, (2015). https://doi.org/10.1145/2733373.2807412
127. X. Cao, D. Wipf, F. Wen, G. Duan, J. Sun, in *2013 IEEE International Conference on Computer Vision*, A practical transfer learning algorithm for face verification, (2013), pp. 3208–3215. https://doi.org/10.1109/ICCV.2013.398
128. D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, in *ECCV 2012*, Bayesian face revisited: a joint formulation, (2012). https://www.microsoft.com/en-us/research/publication/bayesian-face-revisited-a-joint-formulation/
129. D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, in *Computer vision – ECCV 2012*. ed. by A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Bayesian face revisited: a joint formulation (Springer, Berlin, Heidelberg, 2012), pp. 566–579
130. Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, X. Chen, A benchmark and comparative study of video-based face recognition on cox face database. IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. **24** (2015). https://doi.org/10.1109/TIP.2015.2493448
131. J. Chen, T. Takiguchi, Y. Ariki, A robust SVM classification framework using PSM for multi-class recognition. EURASIP J. Image Video Process. **2015**(1), 7 (2015). https://doi.org/10.1186/s13640-015-0061-x
132. V. Vapnik, R. Izmailov, Knowledge transfer in SVM and neural networks. Ann. Math. Artif. Intell. **81**(1), 3–19 (2017). https://doi.org/10.1007/s10472-017-9538-x
133. P. Wei, Z. Zhou, L. Li, J. Jiang, Research on face feature extraction based on k-mean algorithm. EURASIP J. Image Video Process. **2018**(1), 83 (2018). https://doi.org/10.1186/s13640-018-0313-7
134. J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, K. Murphy, Speed/accuracy trade-offs for modern convolutional object detectors. CoRR. **abs/1611.10012** (2016). 1611.10012
135. S. W. Arachchilage, E. Izquierdo, in *2019 IEEE Visual Communications and Image Processing (VCIP)*, A framework for real-time face-recognition, (2019), pp. 1–4. https://doi.org/10.1109/VCIP47243.2019.8965805
136. W. Jiang, W. Wang, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Face detection and recognition for home service robots with end-to-end deep neural networks, (2017), pp. 2232–2236. https://doi.org/10.1109/ICASSP.2017.7952553
137. C. Ding, D. Tao, Robust face recognition via multimodal deep face representation. IEEE Trans. Multimed. **17**(11), 2049–2058 (2015). https://doi.org/10.1109/TMM.2015.2477042
138. Z. Cui, S. Shan, X. Chen, L. Zhang, in *Face and Gesture 2011*, Sparsely encoded local descriptor for face recognition, (2011), pp. 149–154. https://doi.org/10.1109/FG.2011.5771389
139. C. Ding, J. Choi, D. Tao, L. S. Davis, Multi-directional multi-level dual-cross patterns for robust face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 518–531 (2016). https://doi.org/10.1109/TPAMI.2015.2462338
140. Z. Cao, Q. Yin, X. Tang, J. Sun, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Face recognition with learning-based descriptor, (2010), pp. 2707–2714. https://doi.org/10.1109/CVPR.2010.5539992
141. Z. Cui, W. Li, D. Xu, S. Shan, X. Chen, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Fusing robust face region descriptors via multiple metric learning for face recognition in the wild, (2013), pp. 3554–3561. https://doi.org/10.1109/CVPR.2013.456
142. A. Afaneh, F. Noroozi, Ö. Toygar, Recognition of identical twins using fusion of various facial feature extractors. EURASIP J. Image Video Process. **2017**(1), 81 (2017). https://doi.org/10.1186/s13640-017-0231-0
143. B. Wu, S. Lyu, B. Hu, Q. Ji, in *2013 IEEE International Conference on Computer Vision*, Simultaneous clustering and tracklet linking for multi-face tracking in videos, (2013), pp. 2856–2863. https://doi.org/10.1109/ICCV.2013.355
144. S. Zhang, J. Huang, J. Lim, Y. Gong, J. Wang, N. Ahuja, M. Yang, Tracking persons-of-interest via unsupervised representation adaptation. CoRR. **abs/1710.02139** (2017). 1710.02139
145. M. Roth, M. Bäuml, R. Nevatia, R. Stiefelhagen, in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Robust multi-pose face tracking by multi-stage tracklet association, (2012), pp. 1012–1016
146. B. Wu, Y. Zhang, B. Hu, Q. Ji, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Constrained clustering and its application to face clustering in videos, (2013), pp. 3507–3514. https://doi.org/10.1109/CVPR.2013.450
147. C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, Q. Dai, Cross-modality bridging and knowledge transferring for image understanding. IEEE Trans. Multimed. **21**(10), 2675–2685 (2019). https://doi.org/10.1109/TMM.2019.2903448
148. C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, Q. Dai, A fast uyghur text detector for complex background images. IEEE Trans. Multimed. **20**(12), 3389–3398 (2018). https://doi.org/10.1109/TMM.2018.2838320
149. S. Navid Hajimirza, M. Proulx, E. Izquierdo, Reading users' minds from their eyes: a method for implicit image annotation. Multimed. IEEE Trans. **14**, 805–815 (2012). https://doi.org/10.1109/TMM.2012.2186792
150. C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, Stat: spatial-temporal attention mechanism for video captioning. Trans. Multimed. IEEE, 1–1 (2019). https://doi.org/10.1109/TMM.2019.2924576
151. X. Wang, Q. Ruan, Y. Jin, G. An, Three-dimensional face recognition under expression variation. EURASIP J. Image Video Process. **2014**(1), 51 (2014). https://doi.org/10.1186/1687-5281-2014-51

152. L. Yang, J. Ma, J. Lian, Y. Zhang, H. Liu, Deep representation for partially occluded face verification. EURASIP J. Image Video Process. **2018**(1), 143 (2018). https://doi.org/10.1186/s13640-018-0379-2
153. E. Izquierdo, M. Ghanbari, Key components for an advanced segmentation system. IEEE Trans. Multimed. **4**(1), 97–113 (2002). https://doi.org/10.1109/6046.985558
154. H. Jiang, G. Zhang, H. Wang, H. Bao, Spatio-temporal video segmentation of static scenes and its applications. IEEE Trans. Multimed. **17**(1), 3–15 (2015). https://doi.org/10.1109/TMM.2014.2368273
155. X. Sun, J. Foote, D. Kimber, B. S. Manjunath, Region of interest extraction and virtual camera control based on panoramic video capturing. IEEE Trans. Multimed. **7**(5), 981–990 (2005). https://doi.org/10.1109/TMM.2005.854388
156. L. Dong, L. He, M. Mao, G. Kong, X. Wu, Q. Zhang, X. Cao, E. Izquierdo, Cunet: a compact unsupervised network for image classification. IEEE Trans. Multimed. **20**(8), 2012–2021 (2018). https://doi.org/10.1109/TMM.2017.2788205
157. J. Bruna, S. Mallat, Invariant scattering convolution networks. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1872–1886 (2013). https://doi.org/10.1109/TPAMI.2012.230
158. R. Hankins, Y. Peng, H. Yin, in *Intelligent Data Engineering and Automated Learning – IDEAL 2018*. ed. by H. Yin, D. Camacho, P. Novais, and A. J. Tallón-Ballesteros, Towards complex features: competitive receptive fields in unsupervised deep networks (Springer, Cham, 2018), pp. 838–848

Publisher's Note