Deep Learning for Understanding Faces

Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun-Cheng Chen, Vishal M. Patel, Carlos D. Castillo and Rama Chellappa

Abstract

Recent developments in deep convolutional neural networks (DCNNs) have shown impressive performance improvements on various object detection/recognition problems. This has been made possible due to the availability of large annotated data, a better understanding of the non-linear mapping between images and class labels as well as the affordability of powerful GPUs. These developments in deep learning have also improved the capabilities of machines in understanding faces and automatically executing the tasks of face detection, pose estimation, landmark localization, and face recognition from unconstrained images and videos. In this paper, we provide an overview of deep learning methods used for face recognition. We discuss different modules involved in designing an automatic face recognition system and the role of deep learning for each of them. Some open issues regarding DCNNs for face recognition problems are then discussed. The paper should prove valuable to scientists, engineers and end users working in the fields of face recognition, security, visual surveillance, and biometrics.

I. Introduction: What We Can Learn From Faces

Facial analytics is a challenging problem in computer vision and has been actively researched for over two decades [1]. The goal is to extract as much information as possible, such as location, pose, gender, ID, age, emotion, etc from a face. Applications of this technology include detecting and identifying a person of interest from surveillance videos, active authentication of users' cell phones, payment transactions using face biometrics, smart cars, etc. In addition, there has been a growing interest in face recognition and verification from unconstrained images and videos which also involve subtasks, such as face detection, facial landmark localization, etc. Sample outputs of a typical facial analysis system [2] are shown in Figure 9.

In this paper, we present an overview of recent automatic face identification and verification systems based on deep learning. Three modules are typically needed for such an automatic system. First, a face detector is applied to localize faces in images or videos. For a robust system, the face detector should be capable of detecting faces with varying pose, illumination and scale. Also, the locations and sizes of face bounding boxes should be precisely determined so that they contain minimal amount of background. Second, a fiducial point detector is used to localize the important facial landmarks such as eye centers, nose tip, mouth corners, etc. These points are used to align the faces to normalized canonical coordinates to mitigate the effects of in-plane rotation and scaling. Third, a feature descriptor that encodes the identity information is extracted from the aligned face. Given the face representations, similarity

Rama Chellappa, Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal and Navaneeth Bodla are with the department of Electrical and Computer Engineering at University of Maryland College Park, MD USA {rama, rranjan1, swamiviv, ankan, nbodla}@umiacs.umd.edu.

Jun-Cheng Chen and Carlos D. Castillo are with the University of Maryland Institute for Advanced Computer Studies (UMIACS), College Park, MD USA {pullpull, carlos}@umiacs.umd.edu.

Vishal M. Patel is with the department of Electrical and Computer Engineering at Rutgers University, Piscataway, NJ USA vishal.m.patel@rutgers.edu.

scores are then obtained between them using a metric. If this metric is lower than a threshold, it signifies that the two faces are from the same subject. Since the early nineties, many face identification/verification algorithms have been shown to work well on images and videos that are collected in controlled settings. However, the performance of these algorithms often degrades significantly on face images or videos that have large variations in pose, illumination, resolution, expression, aging, background clutter and occlusion (see Figure 6). Moreover, for the application of video surveillance where a subject needs to be identified from hundreds of low resolution videos, the algorithm should be fast and robust.

To overcome these challenges, researchers have employed deep learning for computing the features required for facial analysis. Deep convolutional neural networks (DCNNs) have been shown to be effective for image analysis tasks from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 [3]. Over the last five years, DCNNs have been used for solving numerous computer vision problems such as object recognition [3]–[5], object detection [6]–[8], etc. A typical DCNN is a hierarchy of convolutional layers with Rectified Linear Unit (ReLU) activation function, consisting of millions of trainable parameters. Owing to the learned rich and discriminative representations, DCNNs have recently been successfully applied for face detection [2], [9], [10], fiducial point localization [2], [10], [11] and face recognition and verification [12]. A key contributing factor to this exceptional performance is the availability of large amount of annotated unconstrained face datasets like CASIA-WebFace [13], MegaFace [14], [15], MS-Celeb-1M [16], and VGGFace [17] for the face recognition task, and WIDER FACE [18] for the face detection task. The large amount of training data represents significant variations in pose, illumination, expression and occlusion which enables the DCNNs to be robust to these variations and extract meaningful features required for the task.

The main goal of this paper is to present an overview of recently developed DCNN-based methods for designing an automatic face recognition system, discuss their advantages and disadvantages, and identify interesting open problems. The rest of the paper is organized as follows. In the following sections, we will first discuss recent developments in: (1) face detection (Section II), (2) fiducial point detection (Section III), (3) face recognition (Section IV), (4) facial attributes for face recognition (Section V). In (Section VI), we will discuss various multi-task learning approaches for face analytics. Finally, in Section VII, we will discuss some open issues in designing a robust face recognition system using deep features and conclude in Section VIII.

II. Face Detection in Unconstrained Images

Face detection plays a crucial role in a face recognition pipeline and forms the first step in an automatic face recognition system. Given an input image, a face detector needs to detect all the faces in the image and return bounding-box coordinates for each of them. The key challenge in unconstrained face detection is that traditional features like Haar wavelets and histogram of oriented gradient (HOG) do not capture salient facial information at different resolution, view-point, illumination, expression, skin color, occlusions, and cosmetics conditions. The limitation is more due to the features used than the classifiers. However, with recent advances in deep learning techniques and the availability of GPUs, it is becoming possible to use DCNNs for feature extraction. It has been shown in [3] that a DCNN pretrained with a large generic dataset can be used as a meaningful feature extractor. The

deep features thus obtained have been used extensively for object and face detection. The DCNN-based methods for face detection can be divided into two sub-categories: 1) region-based approach, and 2) sliding window approach.

A. Region-based

The region-based approach generates a pool of generic object-proposals (around 2000 per image), and a DCNN is used to classify whether a given proposal contains a face or not. Most DCNN-based approaches are proposal-based [2], [10], [19]. A general pipeline for region-based face detectors consists of off-the-shelf object proposal generators like Selective Search [20], followed by a DCNN that classifies these proposals as face/non-face. HyperFace [10] and All-In-One Face [2] follow this pipeline for face detection.

Faster R-CNN: A more recent face detector using Faster R-CNN [19] uses a common DCNN to generate the proposals and classify faces. It can also simultaneously regress the bounding-box coordinates for each face proposal. Li *et al.* [21] proposed a multi-task face detector using the framework of faster R-CNN which integrates a DCNN and a 3D mean face model. The 3D mean face model is used to improve the performance of fiducial detection of region proposal network (RPN) which greatly helps face proposal pruning and refinement after face normalization. Similarly, Chen *et al.* [22] proposed a face detector by training a multi-task RPN for face and fiducial detections to generate high quality face proposals while reducing the redundant ones to strike a balance between high recall and precision. The face proposals are then normalized by the detected fiducial points and refined using a DCNN face classifier for improved performance. Recently, Najibi *et al.* proposed the Single Stage Headless face detector (SSH) [23] which uses RPN on VGG-16 [24] network without the fully connected layers.

A major drawback of region-based face detectors is that difficult faces are hard to be captured in any object proposal which results in a low recall. Also, the additional proposal generation task increases the overall computation time.

B. Sliding Window-based

The sliding window-based method computes a face detection score and bounding box co-ordinates at every location in a feature map at a given scale. This procedure is fast compared to region-based methods and can be implemented using just a convolution operation which works in a sliding window fashion. Detections at different scales are typically carried out by creating an image pyramid at multiple scales. DCNN-based face detectors built using the sliding window technique include DP2MFD [9] and DDFD [25]. Faceness [26] produces face-part responses along with full face response, and combines them based on their spatial configuration to determine the final face score. Li *et al.* [27] proposed a cascade architecture based on DCNNs which operates at multiple resolutions, quickly rejects background objects at low-resolution stages, and only evaluates a small number of challenging candidates at high-resolution stages.

Single Shot Detector: Liu *et al.* [8] proposed the SSD (Single Shot Detector) for the task of object detection. SSD is a sliding window-based detector, but instead of creating image pyramid at different scales, it utilizes the inbuilt pyramid structure present in DCNNs. Features are pooled from the intermediate layers of DCNN at different scales

which are used for object classification and bounding-box regression. Since, the detections are obtained in a single forward pass of the network, the overall computation time of SSD is lower than Faster R-CNN [19]. Few SSD-based face detectors have been proposed recently. Yang *et al.* proposed ScaleFace [28] to detect faces of different scales from multiple layers of the network and fuse them in the end. Zhang *et al.* proposed S3FD [29] which uses a scale-equitable framework and scale compensation anchor matching strategy for improved detection of small faces.

Figure. 1 shows the architectures for the above-mentioned approaches.

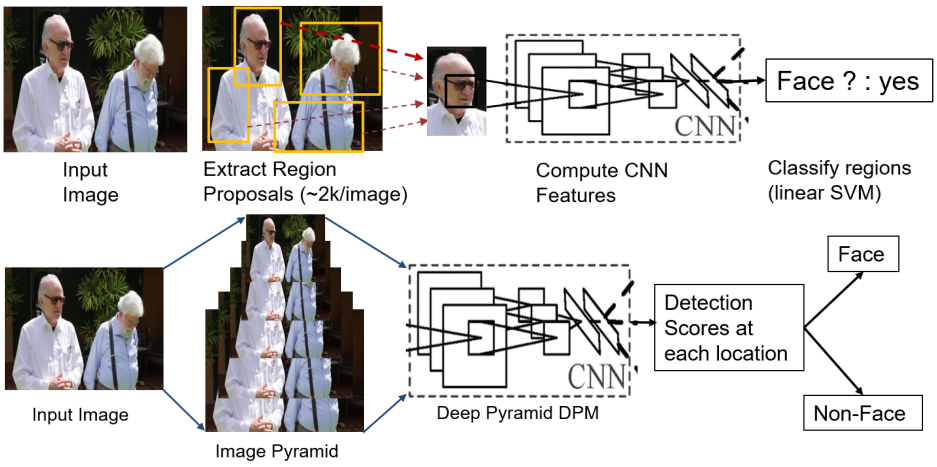


Fig. 1. A typical DCNN-based face detector using region-proposal approach (top) and sliding window approach (bottom).

The availability of large training datasets for unconstrained face detection has also contributed significantly to the success of DCNN-based face detectors. FDDB [30] dataset is the most widely used benchmark for unconstrained face detection. It consists of $2,845$ images containing a total of $5,171$ faces collected from news articles on the Yahoo website. MALF [31] dataset consists of $5,250$ high-resolution images containing a total of $11,931$ faces. The images were collected from Flickr and image search service provided by Baidu Inc. These datasets contain faces in unconstrained settings with large variations in occlusion, pose and illumination.

The WIDER Face [18] dataset contains a total of $32,203$ images, with 50% samples for training and 10% for validation. This dataset contains faces with large variations in pose, illumination, occlusion and scale. Face detectors trained on this dataset have achieved improved performance [19], [23], [28], [29], [32], [33]. The evaluation results for this dataset reveal that finding small faces in the crowd is still challenging. A recent method proposed by Hu *et al.* [33] shows that contextual information is necessary for detecting tiny faces. It captures semantics from lower level features and context from higher level features of a DCNN to detect tiny faces (see Figure 2). Due to space limitations, we are unable to discuss more traditional face detection methods. We refer the interested readers to [34] for more detailed discussions of traditional cascade-based and DPM-based methods. In addition, for videos with multiple faces, one also needs to perform face associations to extract face tracks of each subject. We also refer the readers to [12] for a discussion on the role of face association in video-based face recognition. Fig. 3(a) provides a comparative performance evaluation of different face detectors on the FDDB [30] dataset.

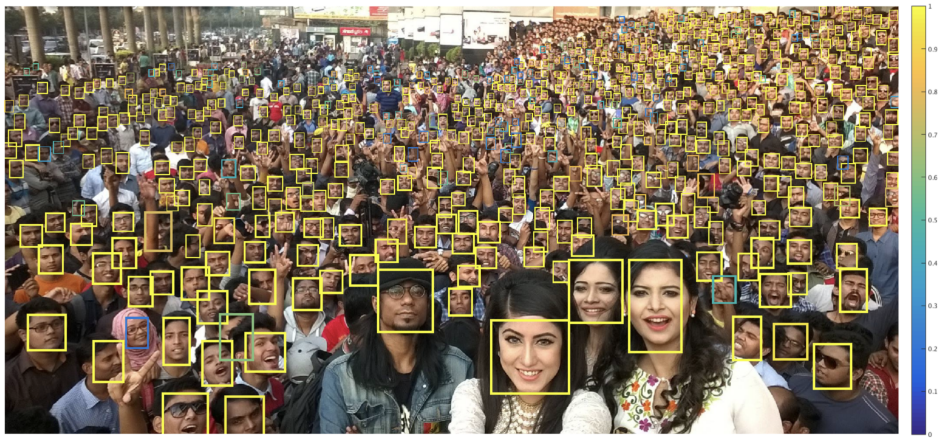
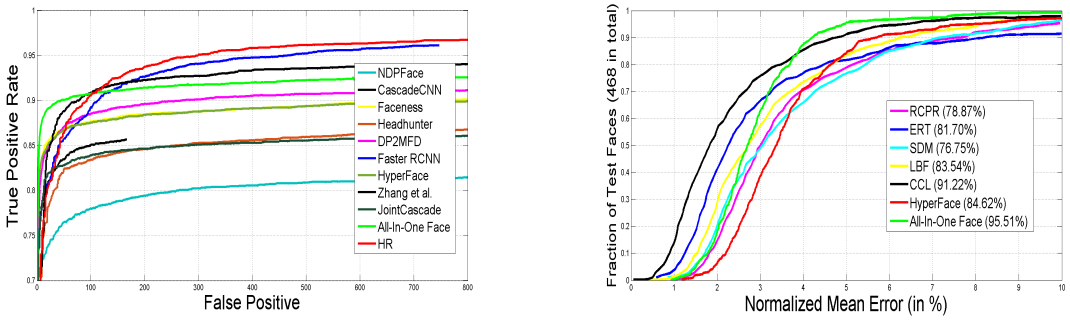


Fig. 2. HR [33] face detector output. The algorithm is able to detect very small faces by making use of novel characterizations of scale, resolution, and context. Detector confidence is given by the color bar on the right.

III. Finding Crucial Facial Keypoints and Head Orientation

Facial keypoint detection is also an important pre-processing component for face recognition and verification tasks. Facial keypoints such as eye centers, nose tip, mouth corners, etc. can be used to align the face into canonical coordinates, and such face normalization helps face recognition [35] and attribute detection. Head-pose estimation is required for pose-based face analysis. Both these problems have been researched extensively over the last few years. Most of the existing facial keypoint localization algorithms use either a model-based approach or a cascaded regression-based approach. Wang *et al.* [36] provides a comprehensive survey of traditional model-based methods, including active appearance model (AAM), active shape model (ASM), constrained local model (CLM), and some regression methods, such as supervised descent method (SDM). Chrysos *et al.* [37] also summarizes fiducial point tracking across videos using traditional fiducial detection methods. We refer the interested readers to these works for more details. In this paper, we summarize recently developed fiducial detection methods based on DCNNs.



(a) (b)

Fig. 3. (a) ROC curves for different face detection methods on FDDB dataset. (b) Landmarks localization evaluation on AFW dataset using the protocol described in [38] for recently published methods such as CCL [38], HyperFace [10], All-In-One Face [2], LBF [39], SDM [40], ERT [41] and RCPR [42]. The numbers in the legend are the fraction of faces having normalized mean error less than 5%.

A. Model-based

Model-based approaches, such as AAM, ASM, and CLM, learn a shape model during training and use it to fit new faces during testing. Recently, Antonakos *et al.* [43] proposed a method by modeling the appearance of the object using multiple graph-based pairwise normal distributions (Gaussian Markov Random Field) between patches extracted from the regions. However, the learned models lack the power to capture complex face image variations in pose, expression and illumination. Also, they are sensitive to the initialization in gradient descent optimization. Besides 2D shape, face alignment methods based on 3D models have been recently developed. PIFA [44] by Jourabloo *et al.* suggested a 3D approach that employed cascaded regression to predict the coefficients of 3D to 2D projection matrix and the base shape coefficients. Another recent work from Jourabloo *et al.* [45] formulated the face alignment problem as a dense 3D model fitting problem, where the camera projection matrix and 3D shape parameters were estimated by a cascade of DCNN-based regressors. 3DDFA [46] by Zhu *et al.* fitted a dense 3D face model to the image via DCNN, where the depth data is modeled in a Z-Buffer.

B. Cascaded Regression-based

Since face alignment is naturally a regression problem, a multitude of regression-based approaches has been proposed in recent years. In general, these methods learn a model that directly maps the image appearance to the target output. Nevertheless, the performance of these methods depends on the robustness of local descriptors. Sun *et al.* [47] proposed a cascade of carefully designed DCNNs, in which at each level, outputs of multiple networks are fused for landmark estimation and achieve good performance. Zhang *et al.* [48] proposed a coarse-to-fine auto-encoder networks approach, which cascades several successive stacked auto-encoder networks (SAN). The first few SANs predict the rough location of each facial landmark, and subsequent SANs progressively refine the landmarks by taking as input the local features extracted around the detected landmarks at the current stage with higher resolution. Kumar *et al.* [11] used a single DCNN, carefully designed to provide a unique key-point descriptor and achieved better performance (see Fig. 4). Xiong *et al.* [49] suggested domain dependent descent maps. Zhu *et al.* [38] observed that optimizing the base shape coefficients and projection is indirect and sub-optimal since smaller parameter errors are not necessarily equivalent to smaller alignment errors. Therefore, they proposed cascade compositional learning (CCL) [38] to develop head pose-based and domain selective regressors by partitioning the optimization domain into multiple directions based on head pose and combining the results of multiple domain regressors through composition estimator function. Trigeorgis *et al.* [50] proposed a combined and jointly trained convolutional recurrent neural network architecture that enables end-to-end training of the regressors jointly used in the cascaded regression framework. It avoids training each regressor independently and cancelling of the descent direction of the regressors. Bulat *et al.* [51] proposed a DCNN architecture which first performs facial part detection by using a score map generated from the DCNN features of the first few layers to roughly localize each facial landmark and then followed by a regression branch to refine the detection results. Therefore, the algorithm is not sensitive to the quality of detected face bounding boxes, and the system can be trained end-to-end. Kumar *et al.* [52] also proposed an effective iterative method for keypoint estimation and pose prediction of unconstrained faces by

learning efficient heatmap-based DCNN regressors for the face alignment problem where the heatmaps provide a probability value, indicating the existence of a certain joint at a specific location.

On the other hand, since different datasets provide different facial landmark annotations, 300 Faces In-the-Wild database (300W) [53] has been developed as a benchmark for enabling a fair comparison among various fiducial detection methods, and it contains more than 12,000 images annotated with 68 landmarks including LFPW [36], Helen [36], AFW [36], Ibug [36] and 600 test images. (*i.e.*, 300 indoors and 300 outdoors.)

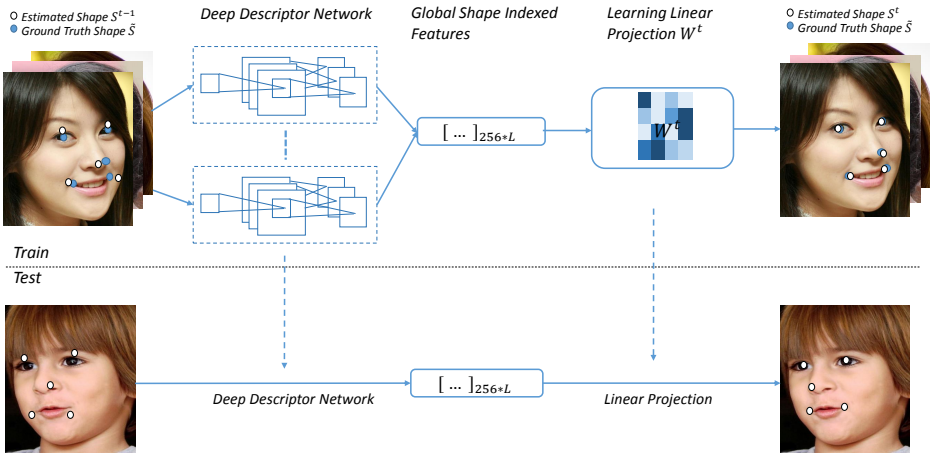


Fig. 4. Model diagram of regression-based keypoint estimation method using DCNNs [11]

Besides using a 2D transformation for face alignment, Hassner *et al.* [54] proposed an effective method to frontalize faces with the help of a generic 3D face model. However, the effectiveness of the method also highly relies on the quality of the detected facial landmarks (i.e., the method usually introduces undesirable artifacts when the quality of facial landmarks is poor). In addition, quite a few methods have been developed for multi-task learning (MTL) approaches for face detection which involve simultaneous training of face detection task along with a correlated task such as facial keypoint estimation. MTL helps the network to build a synergy and learn robust features since the network is aided by additional supervision. For instance, information about eye centers, nose tip, etc. obtained from the keypoints can help the network determine the structure of the face. Zhang *et al.* [32], Chen *et al.* [22], Li *et al.* [21], and HyperFace [10] train the face detection network along with the keypoint estimation task, and All-In-One Face [2] extends the MTL approach to additional tasks such as face verification, gender, smile and age estimation. Section VI discusses recent advances in MTL for face analytics. Fig. 3(b) provides a comparative performance evaluation of different algorithms for the task of facial keypoints estimation on the AFW [55] dataset.

IV. Face Identification and Verification

In this Section, we review the relevant works on face identification and verification. In Figure 5, we illustrate a general pipeline for training and testing a face identification/verification system using deep convolutional neural networks. There are two major components for face identification/verification: (1) robust face representation and (2) discriminative classification model (face identification) or similarity measure (face verification). Since we focus on

deep learning-based approaches, we refer the interested readers to [56] which summarizes traditional approaches, including local binary pattern (LBP) and Fisher vector (FV), and metric learning, such as one-shot similarity kernel (OSS), Mahalanobis metric learning, cosine metric learning, large margin nearest neighbor (LMNN), attribute-based classifier, and joint Bayesian (JB).

A. Robust Feature Learning for Faces using Deep Learning

Learning invariant and discriminative feature representations is a critical step in a face recognition system. Deep learning methods have shown that compact and discriminative representations can be learned using a DCNN trained with very large datasets. We first review some important face recognition works using deep learning followed by a brief summary of recent methods for feature representation learning.

Instead of relying on hand-crafted features such as the popular face descriptor LBP, Huang *et al.* [57] proposed to learn the face representation using convolutional deep belief networks based on local convolutional restricted Boltzmann machines which first learns useful representation in an unsupervised manner from an unlabeled natural image dataset containing images of natural scenes and then transfers the learned representation to face identification/verification task through classification models (*e.g.* SVM) and metric learning approaches (*e.g.* OSS). The approach achieved satisfactory results on the LFW dataset without using large-scale annotated face datasets.

One of the earlier large scale applications of 3D model based alignment along with DCNNs for face recognition was proposed by Taigman *et al.* in their DeepFace [58] approach. They derive a face representation using a nine-layer deep neural network, that involves more than 120 million parameters and uses several locally connected layers without weight sharing, rather than the standard convolutional layers. A proprietary face dataset that consists of four million facial images belonging to more than 4,000 identities was used to train their system.

Since collecting a large-scale annotated dataset is costly, Sun *et al.* proposed DeepID frameworks [59]–[61] with JB for face verification which utilized an ensemble of shallower and smaller deep convolutional networks (*i.e.* each DCNN consists of four convolutional layers and use $39\times31\times1$ patches as the input) than DeepFace and were trained using 202,599 images of 10,177 subjects. The large number of distinct identities of the dataset and an ensemble of DCNNs for different local and global face patches make DeepID learn discriminative and informative face representation. The approach was the first to achieve results that surpass human performance for face verification on the LFW dataset.

Schroff *et al.* proposed a CNN-based approach for face recognition called FaceNet [62] which directly optimizes the embedding itself, rather than an intermediate bottleneck layer as in other deep learning approaches. In order to train, they use triplets of roughly aligned matching / non-matching face patches generated using an online triplet mining method. Their approach was trained on a large proprietary face dataset consisting of 100M-200M face thumbnails consisting of about 8M different identities.

Yang *et al.* [13] collected a public large-scale annotated face dataset, CASIA-WebFace, which consists of 494,414 face images for 10,575 subjects collected from the IMDB website. They trained a DCNN with 5 million parameters. The model followed with JB achieved satisfactory results on the LFW dataset. CASIA-WebFace is widely used to train various DCNN models in the face recognition community.

Parkhi *et al.* [17] also collected a public large-scale annotated face dataset, VGGFace, which consists of 2.6 million face images for 2.6 thousand subjects, and trained a DCNN based on the well-known VGGNet [24] for object recognition, followed by triplet embedding for face verification. The trained DCNN model using VGGFace achieves comparable results on both still-face (*i.e.* LFW) and video-face (*i.e.* YTF) datasets with other state-of-the-art approaches using only a single model and a publicly available dataset. This demonstrates the usefulness of the VGGFace dataset to face recognition community.

In a recent work, AbdAlmageed *et al.* [63] handle pose variations by training separate DCNN models for frontal, half-profile and full-profile poses in order to improve face recognition performance in the wild. Furthermore, Masi *et al.* [64] utilized 3D morphable models to augment the CASIA-WebFace dataset with large amounts of synthetic faces to improve the recognition performance instead of collecting more data through crowdsourcing the annotation tasks. Ding *et. al.* [65] proposed to fuse the deep features around facial landmarks from different layers followed by a new triplet loss function which achieves state-of-the-art performance for video-based face recognition. Wen *et. al.* [66] proposed a new loss function which takes the centroid for each class into consideration and uses it as a regularization constraint to the softmax loss in addition to a residual neural network for learning more discriminative face representation. Liu *et al.* [67] proposed a novel angular loss based on the modified softmax loss. It results in a discriminative angular feature representation for faces which is optimized to the commonly used similarity metric, cosine distance, and the model achieves recognition performance comparable to state-of-the-art methods that use much smaller training dataset. Ranjan *et al.* [68] also trained with softmax loss regularized with scaled L_2-norm constraint on a subset of recently released MS-Celeb-1M face dataset. They show that the regularized loss optimizes the angular margin between classes. The method achieves state-of-the-art results on the IJB-A dataset [69]. Besides the commonly-used average aggregation for per-frame video face representation, Yang *et al.* proposed a neural aggregated network (NAN) [70] to perform dynamically weighted aggregation on the features from multiple face images or face frames of a face video to yield a succinct and robust representation for video face recognition. The approach demonstrated state-of-the-art results on multiple image-set and video face datasets. Bodla *et al.* [71] proposed a fusion network to combine face representations from two different DCNN models to improve the recognition performance.

B. Discriminative Metric Learning for Faces

Learning a classifier or a similarity measure from data is another key component for improving the performance of a face recognition system. Many approaches have been proposed in the literature that essentially exploit the label information from face images or face pairs. Hu *et al.* [72] learned a discriminative metric within the deep neural network framework. Schroff *et al.* [62] and Parkhi *et al.* [17] optimized the DCNN parameters based on the triplet loss which directly embeds the DCNN features into a discriminative subspace and presented promising results for face verification. In [73], a discriminant low-rank embedding is learned using a probabilistic model for face verification and clustering. Song *et al.* [74] proposed a method to fully utilize the training data in a batch by considering the full pairwise distances between samples.

Besides supervised learning of DCNN for face recognition, Yang *et al.* [75] proposed learning deep representations and image clusters jointly in a recurrent framework. Each image is treated as separate clusters at the beginning, and a deep network is trained using this initial grouping. Deep representation and cluster members are then iteratively refined until the number of clusters reached the predefined value. The unsupervised learned representation has demonstrated promising results on various tasks, including face recognition, digit classification, etc. Zhang *et al.* [76] proposed to cluster face images in videos by alternating between deep representation adaption and clustering. Trigeorgis *et al.* [77] proposed a deep semi-supervised non-negative matrix factorization to learn hidden representations that allow themselves to an interpretation of clustering according to different, unknown attributes of a given face dataset, such as pose, emotion, and identity. Their method also demonstrates promising results on challenging face datasets. On the other hand, Lin *et al.* [78] proposed an unsupervised clustering algorithm that exploits the neighborhood structure between samples and implicitly performs domain adaptation to achieve improved clustering performance. They also demonstrated an application of the clustering algorithm to curate a large-scale noisy face dataset, such as MS-Celeb-1M [79].

C. Implementation

Face Recognition can be sub-divided into two tasks: 1) Face Verification, and 2) Face Identification. For the task of face verification, given a pair of face images, the system needs to predict whether they belong to the same person or not. For the task of face identification, given a face image with unknown identity, the system should determine the subject's identity by matching the features from the database.

For both these tasks, obtaining discriminative and robust identity features are important. For the task of face verification, the faces are first localized by the face detector and normalized to the canonical coordinate through similarity transform using the detected fiducial points. Then, each of the face images is passed through a DCNN to obtain a feature representation. Once the features are obtained, a score measure is computed based on a similarity metric. Most commonly used similarity metrics are: a) L_2 distance between the face features, b) $CosineSimilarity$ which provides a measure of how close the features are in the angular space. One can also use multiple DCNNs for improved performance by fusing the network features or the similarity scores such as DeepID framework [59]–[61] or the fusion network [71] approach. For the task of face identification, the face images in the gallery are passed through DCNN and the features for each identity are stored in the database. When a new face image is provided to the system, it computes its feature representation, and a similarity score is obtained for each of the identities in the gallery.

D. Training Datasets for Face Recognition

In Table I, we summarize the public datasets used for testing the algorithm performance and for training the DCNN models on face recognition tasks. MS-Celeb-1M [79] is the current largest public face recognition dataset, which contains over 10 million labeled face images of top 100,000 distinct identities from the 1 million celebrity list with significant pose, illumination, occlusion, and other variations. Since this dataset also contains large amount

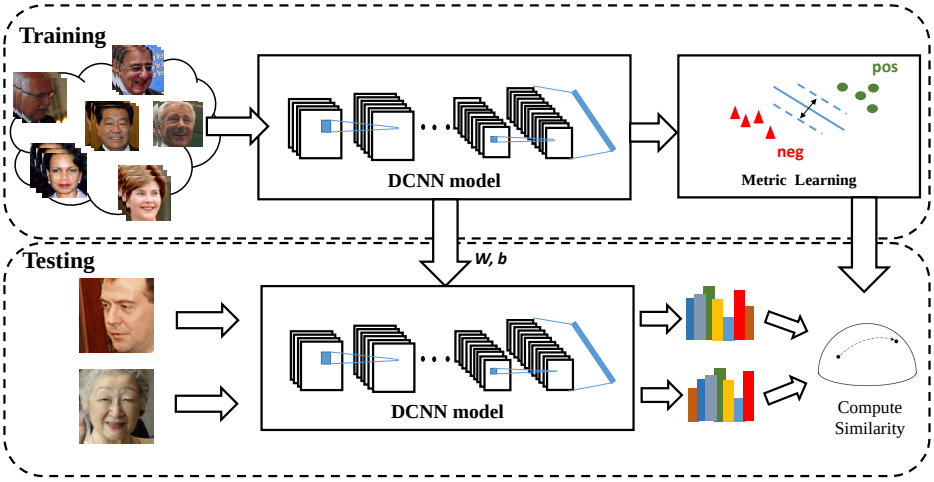


Fig. 5. A general pipeline for training and testing a face verification system using DCNN.

of label noise, interested readers are referred to [78] for an example of dataset curation by applying subject clustering. For other datasets, the CelebA dataset [80] is annotated with forty face attributes and five key points by a professional labeling company for 202,599 face images of 10,000 subjects. The dataset can be used for training face recognition, face attribute classification, and fiducial prediction models. The CASIA-WebFace [13] is also one of popular public large-scale annotated datasets for face recognition which consists of 494,414 face images for 10,575 subjects collected from the IMDB website. The VGGFace [17] is another public dataset for training face recognition algorithms and consists of 2.6 million face images for 2.6 thousand subjects. The MegaFace [14], [15] is the dataset used to test the robustness of face recognition algorithms in the open-set setting with 1 million distractors. There are two parts for the dataset where the first one allows to use any external training datasets, and the other provides 4.7 million face images of 672 thousand subjects. The LFW [81] dataset contains 13,233 face images of 5,749 subjects from the Internet where 1,680 subjects have two or more images. The dataset is mainly used to evaluate the performance of the still face recognition algorithms and most of the faces are frontal. The IJB-A [69] dataset contains 500 subjects with 5,397 images and 2,042 videos splitting into 20,412 frames. The dataset is designed to test the robustness of the algorithms under large pose, illumination, and image/video frame quality variations in the mixed-media setting (*i.e.* each gallery and probe set consists of one or more images and video frames). The YTF [82] dataset contains 3,425 videos of 1,595 different subject and is the standard dataset used to evaluate video-face recognition algorithms. The PaSC [83] dataset consists of 2,802 videos of 293 subjects and is used to test the performances of video-face algorithms under large pose, illumination, blur variations where the videos are captured in controlled and point-and-shoot manners. The Celebrities in Frontal-Profile (CFP) dataset [84] contains 7,000 images of 500 subjects, and is used for evaluating face verification approaches on extreme pose variations. The UMDFaces [85] and UMDFace Video [35] are the datasets which contain 367,888 still images of 8,277 subjects and 22,075 videos of 3,107 subjects. The datasets can be used to training still and video face recognition algorithms. The subjects in UMDFace Video are also in UMDFaces, and this is useful to adapt the

models trained in still-image face datasets to the video domain.

TABLE I
RECENT DATASETS FOR FACE RECOGNITION.

Face Recognition		
Name	#faces	#subjects
MS-Celeb-1M [16]	10M	100K
CelebA [80]	202,599	10,177
CASIA-WebFace [13]	494,414	10,575
VGGFace [17]	2.6M	2,622
Megaface [14], [15]	4.7M	672K
LFW [81]	13,233	5749
IJB-A [69]	25,809	500
YTF [82]	3,425 videos	1,595
PaSC [83]	2,802 videos	293
CFP [84]	7,000	500
UMDFaces [85]	367,888	8,277
UMDFace Video [35]	22,075 videos	3,107

Recently, Bansal *et al.* [35] studied various features of a good large-scale dataset for face recognition, including (1) Can we train on still images and expect the systems to work on videos? (2) Are deeper datasets better than wider datasets where given a set of images deeper datasets mean more images per subject, and wider datasets mean more subjects? (3) Does adding label noise always leads to improvement in performance of deep networks? (4) Is alignment needed for face recognition? They used CASIA-WebFace [13], UMDFaces [85] and its video extension [35], Youtube Face [82], and the IJB-A dataset [69]. They [35] found that DCNN models trained on a combination of still images and video frames perform better than those trained on only still images or on only video frames. Based on the experiments, it appears that for smaller models, training using wider datasets is better than training using deeper datasets, while for deeper models, training using wider datasets is better. The authors showed that label noise usually hurts the performance of face recognition. It was also found that good alignment can greatly help the face recognition performance.

E. Performance Summary

To sum up, we summarize the performance results of recent face identification and verification algorithms for the well-known Labeled Faces in the Wild (LFW) [81] and IARPA Benchmark A (IJB-A) [69] datasets.

1) Labeled Faces in the Wild: The LFW dataset contains 13,233 face images of 5,749 subjects where 1,680 subjects have two or more images. We show the performance comparisons using the standard protocol of face verification which defines 3,000 positive pairs and 3,000 negative pairs in total and further splits them into 10 disjoint subsets for cross validation. Each subset contains 300 positive and 300 negative pairs. It contains 7,701 images of 4,281 subjects. We show the mean accuracy of state-of-the-art deep learning-based methods: DeepFace

[58], DeepID2 [61], DeepID3 [86], FaceNet [62], Yi *et al.* [13], Wang *et al.* [87], Ding *et al.* [88], Parkhi *et al.* [17], Wen *et al.* [66], Liu *et al.* [67], Ranjan *et al.* [68] and human performance on the "funneled" LFW images.

Method	#Net	Training Set	Metric	Mean Accuracy ± Std
DeepFace [58]	1	4.4 million images of 4,030 subjects, private	cosine	95.92% ± 0.29%
DeepFace	7	4.4 million images of 4,030 subjects, private	unrestricted, SVM	97.35% ± 0.25%
DeepID2 [61]	1	202,595 images of 10,117 subjects, private	unrestricted, Joint-Bayes	95.43%
DeepID2	25	202,595 images of 10,117 subjects, private	unrestricted, Joint-Bayes	99.15% ± 0.15%
DeepID3 [86]	50	202,595 images of 10,117 subjects, private	unrestricted, Joint-Bayes	99.53% ± 0.10%
FaceNet [62]	1	260 million images of 8 million subjects, private	L2	99.63% ± 0.09%
Yi *et al.* [13]	1	494,414 images of 10,575 subjects, public	cosine	96.13% ± 0.30%
Yi *et al.*	1	494,414 images of 10,575 subjects, public	unrestricted, Joint-Bayes	97.73% ± 0.31%
Wang *et al.* [87]	1	494,414 images of 10,575 subjects, public	cosine	96.95% ± 1.02%
Wang *et al.*	7	494,414 images of 10,575 subjects, public	cosine	97.52% ± 0.76%
Wang *et al.*	1	494,414 images of 10,575 subjects, public	unrestricted, Joint-Bayes	97.45% ± 0.99%
Wang *et al.*	7	494,414 images of 10,575 subjects, public	unrestricted, Joint-Bayes	98.23% ± 0.68%
Ding *et al.* [88]	8	471,592 images of 9,000 subjects, public	unrestricted, Joint-Bayes	99.02% ± 0.19%
Parkhi *et al.* [17]	1	2.6 million images of 2,622 subjects, public	unrestricted, TDE	98.95%
Wen *et al.* [66]	1	0.7 million images of 17,189 subjects, public	cosine	99.28%
Liu *et al.* [67]	1	494,414 images of 10,575 subjects, public	cosine	99.42%
Ranjan *et al.* [68]	1	3.7 million images of 58,207 subjects, public	cosine	99.78%
Human, funneled [87]	N/A	N/A	N/A	99.20%

TABLE II

ACCURACY OF DIFFERENT VERIFICATION METHODS ON THE LFW DATASET.

2) IARPA Benchmark A dataset: Besides the LFW dataset, we also present the performance comparisons for the challenging IJB-A dataset. The IJB-A dataset contains 500 subjects with 5,397 images and 2,042 videos splitting into 20,412 frames. Sample images and video frames from the datasets are shown in Figure 6. The ROC curve measures the performance in the verification scenarios, and the CMC score measures the accuracy in closed set identification scenarios. In addition, the IJB-A evaluation protocol consists of face verification (1:1 matching) over 10 splits. Each split contains around 11,748 pairs of templates (1,756 positive and 9,992 negative pairs) on average. Similarly, the face identification (1:N search) protocol also consists of 10 splits, which are used to evaluate the search performance. In each search split, there are about 112 gallery templates and 1,763 probe templates (*i.e.* 1,187 genuine probe templates and 576 impostor probe templates). The training set contains 333 subjects, and the test set contains 167 subjects without any overlapping subjects. Ten random splits of training and testing are provided. Unlike LFW [81] and YTF [89] datasets, which only use a sparse set of negative pairs to evaluate the verification performance, the IJB-A dataset divides the images/video frames into gallery and probe sets so that all the available positive and negative pairs are used for the evaluation. Also, each gallery and probe set consist of multiple templates. Each template contains a combination of images or frames sampled from multiple image sets or videos of a subject. In contrast to LFW and YTF datasets, which only include faces detected by the Viola Jones face detector [90], the images in the IJB-A contain extreme pose, illumination, and expression variations. These factors essentially make the IJB-A a challenging face recognition dataset.

The cumulative match characteristic (CMC) scores and the receiver operating characteristic curves (ROC) are used

Fig. 6. The IJB-A dataset contains faces in large variations of pose, illumination, image quality, occlusion, etc.

to evaluate the performance of different algorithms for face identification/verification. The identification/verification results for the IJB-A dataset are shown in Table III. In addition to using the average feature representation, we also perform media averaging which is to first average the features from the same media (image or video) and then further average the media average features to generate the final feature representation followed by Triplet Probabilistic Embedding proposed in [73].

TABLE III

RESULTS ON THE IJB-A DATASET. THE TAR OF ALL THE APPROACHES AT FAR=0.1, 0.01, AND 0.001 FOR THE ROC CURVES (IJB-A 1:1 VERIFICATION). THE RANK-1 AND RANK-10 RETRIEVAL ACCURACIES OF THE CMC CURVES (IJB-A 1:N IDENTFICATION). WE REPORT AVERAGE AND STANDARD DEVIATION OF THE 10 SPLITS FOR ALL-IN-ONE FACE [2], DCNN$_{all}$, AND AFTER TRIPLET PROBABILISTIC EMBEDDING, DCNN$_{all+tpe}$. ALL THE PERFORMANCE RESULTS REPORTED IN [91], DCNN$_{casia}$ [87], JANUS B (JANUSB-092015), DCNN$_{bl}$ [92], DCNN$_{fusion}$ [12], DCNN$_{3d}$ [64], NAN [70], DCNN$_{pose}$ [63], DCNN$_{tpe}$ [73], TP [93], AND DCNN$_{l2+tpe}$ [68] WHERE RX101 MEANS RESNEXT-101 [94]

IJB-A-Verif	DCNN$_{casia}$ [87]	JanusB [91]	DCNN$_{pose}$ [63]	DCNN$_{bl}$ [92]	NAN [70]	DCNN$_{3d}$ [64]
FAR=1e-3	0.514	0.65	-	-	0.881	0.725
FAR=1e-2	0.732	0.826	0.787	-	0.941	0.886
FAR=1e-1	0.895	0.932	0.911	-	0.978	-
IJB-A-Ident	DCNN$_{casia}$ [87]	JanusB [91]	DCNN$_{pose}$ [63]	DCNN$_{bl}$ [92]	NAN [70]	DCNN$_{3d}$ [64]
Rank-1	0.820	0.87	0.846	0.895	0.958	0.906
Rank-10	-	0.95	0.947	-	0.986	0.977
IJB-A-Verif	DCNN$_{fusion}$ [12]	DCNN$_{tpe}$ [73]	DCNN$_{all}$ [2]	DCNN$_{all+tpe}$ [2]	TP [93]	DCNN$_{l2+tpe}$(RX101) [68]
FAR=1e-3	0.76	0.813	0.787	0.823	-	**0.943**
FAR=1e-2	0.889	0.9	0.893	0.922	0.939	**0.970**
FAR=1e-1	0.968	0.964	0.968	0.976	-	**0.984**
IJB-A-Ident	DCNN$_{fusion}$ [12]	DCNN$_{tpe}$ [73]	DCNN$_{all}$ [2]	DCNN$_{all+tpe}$ [2]	TP [93]	DCNN$_{l2+tpe}$(RX101) [68]
Rank-1	0.942	0.932	0.941	0.947	0.928	**0.973**
Rank-10	0.988	0.977	**0.988**	**0.988**	0.986	**0.988**

Table III summarizes the scores (*i.e.*, both ROC and CMC numbers) produced by different face identification/verification methods on the IJB-A dataset. We compare the results with DCNN$_{casia}$ [87], DCNN$_{bl}$ (bilinear CNN [92]), DCNN$_{pose}$ (multi-pose DCNN models [63]), [70], DCNN$_{3d}$ [64], template adaptation (TP) [93], DCNN$_{tpe}$ [73], DCNN$_{all}$ [2](All-in-One Face), DCNN$_{l2+tpe}$ [68], and the one [91] reported recently by NIST. We also

summarize the details of each method in Table IV.

TABLE IV

COMPARISON OF FACE RECOGNITION ALGORITHMS ON THE IJB-A DATASET WHERE OSS-SVM [93] MEANS SVM WITH ONE-SHOT SIMILARITY FRAMEWORK, TSE [12] MEANS TRIPLET SIMILARITY EMBEDDING, TPE [73] MEANS TRIPLET PROBABILISTIC EMBEDDING, OSS-SVM [93] MEANS SVM USING ONE-SHOT SIMILARITY FRAMEWORK, AND ATTENTIONNET MEANS THE ATTENTION NETWORK WHICH IS USED TO COMPUTE THE WEIGHTS OF EACH FACE FOR FEATURE AGGREGATION AND TO SUBSTITUTE THE AVERAGE REPRESENTATION IN ALL THE OTHER COMPARED METHODS.

Method	Base Network	Loss	Training Set	Metric	Rank-1	Rank-10
DCNN$_{casia}$ [87]	CASIANet	Softmax	494,414 images of 10,575 subjects, public [13]	JB	82	-
DCNN$_{pose}$ [63]	VGGNet	Softmax	494,414 images of 10,575 subjects, public [13]	cosine	84.6	94.7
DCNN$_{bl}$ [92]	VGGNet	Softmax	494,414 images of 10,575 subjects, public [13]	SVM	89.5	-
NAN [70]	GoogLeNet(BN) + AttentionNet	Contrastive	3 million images of 50,000 subjects, private [70]	cosine	95.8	98.6
DCNN$_{3d}$ [64]	VGGNet	Softmax	2,472,070 synthesized images of 10,575 subjects using 3D morphable model (494,414 images out of them are natural face images), public [13]	cosine	90.6	97.7
DCNN$_{fusion}$ [12]	AlexNet+CASIANet	Softmax	494,414 images of 10,575 subjects, public [13]	TSE	94.2	98.8
DCNN$_{tpe}$ [73]	AlexNet	Softmax	494,414 images of 10,575 subjects, public [13]	TPE	93.2	97.7
DCNN$_{all+tpe}$ [2]	AlexNet	Multi-task	993,153 images from multiple datasets, (only 494,414 images out of them contain identity labels from 10,575 subjects), public [2]	TPE	94.7	**98.8**
TP [93]	VGGNet	Softmax	2.6 million images of 2,622 subjects, public [17]	OSS-SVM	92.8	98.6
DCNN$_{l2+tpe}$ [68]	ResNet	L$_2$-Softmax	3.7 million images of 58,207 subjects, public [79]	TPE	**97.3**	**98.8**

V. FACIAL ATTRIBUTES

From a single face, we are able to identify facial attributes such as gender, expression, age, skin-tone, etc. Those attributes are very useful for applications like image retrieval, emotion detection, and mobile security. In biometrics literature, facial attributes are also referred to as "soft biometrics" [95]. Kumar *et al.* [56] introduced the concept of attributes as image descriptors for face verification. They used a collection of 65 binary attributes to describe each face image. Berg *et al.* created classfiers for each pair of people in a dataset and then used these classifiers to create features for a face verification classifier [56]. Here, rather than manually identifying attributes, each person was described by their likeness to other people. This is a way of automatically creating a set of attributes without having to exhaustively hand-label attributes on a large dataset. Recently, DCNNs have been used for attribute classification, demonstrating impressive results. Pose Aligned Networks for Deep Attributes (PANDA) achieved state-of-the-art performance by combining part-based models with deep learning to train pose-normalized DCNNs for attribute classification [96]. Focusing on age and gender, [97] applied DCNNs to the Adience dataset. Liu *et al.* used two DCNNs - one for face localization and the other for attribute recognition - and achieved impressive results on the CelebA and LFWA datasets, outperforming PANDA on many attributes [80].

Instead of treating each attribute to be independent, [99] exploits the correlation amongst attributes to improve image ranking and retrieval by using independently trained attribute classifiers and then learning pairwise correlations based on the outputs of these classifiers. Hand *et al.* [100] train a single attribute network which classifies 40

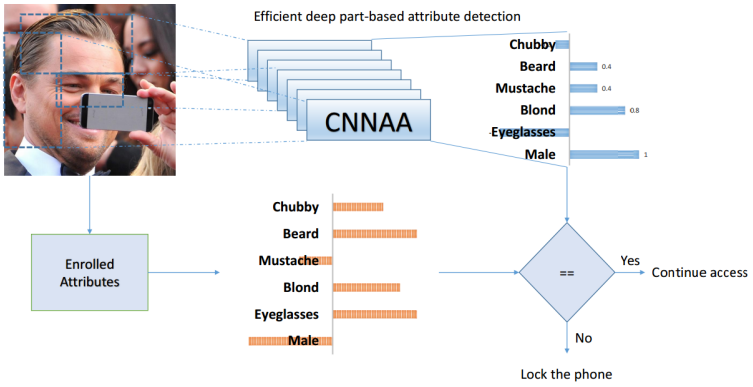


Fig. 7. A sample pipeline of a face attribute detection algorithm for mobile authentication [98].

attributes, sharing information throughout the network, and by learning the relationships among all the 40 attributes, not just attribute pairs. Ranjan *et al.* [2] train a single deep network with multi-task learning which can simultaneously perform face detection, facial landmark detection, face recognition, 3D head pose estimation, gender classification, age estimation, and smile detection. Recently, Gunther *et al.* proposed the AFFACT [101] algorithm that performs alignment-free facial attribute classification. It uses a data augmentation technique that allows a network to classify facial attributes without requiring alignment beyond detected face bounding boxes. The algorithm achieves state-of-the-art results on CelebA [80] dataset with an ensemble of three networks.

In addition, some of the facial attributes can be used [17] to boost the mobile authentication performance. The recent method of Attribute-based Continuous Authentication (ACA) [102], [103] shows that large number of attributes by themselves can give good authentication results on mobile phones. They are semantic features which are easier to learn than facial identities. Also, if they are learned on facial parts, they become less complex. By leveraging these two qualities, Samangouei *et. al.* [98] designed efficient DCNN architectures suitable for mobile devices. The pipeline in Figure 7 illustrates how facial attributes are used for mobile authentication.

VI. Multi-Task Learning for Facial Analysis

In this section, we present the design details of various Multi-Task Learning (MTL) methods for different facial analysis tasks. The MTL framework for machine learning was first analyzed by Caruana [104] in detail. Since then, it has been used for solving multiple problems in Computer Vision. One of the earlier works of MTL-based facial analysis was proposed by Zhu *et al.* [55]. The algorithm was used for solving the tasks of face detection, landmarks localization and head-pose estimation. It used a mixture of trees model with shared pool of parts, where a part represents a landmark location. Another method called JointCascade [105] improved the face detection algorithm by combining the training with the landmarks localization task. Both these methods used hand-crafted features which made it difficult to extend the MTL approach to a wide variety of tasks.

Before the advent of deep learning methods, MTL was restricted to a few sets of tasks because the feature representations required for solving each of the tasks were different. For instance, the task of face detection typically used Histogram of Oriented Gradients (HOG), whereas the task of face recognition used Local Binary Patterns

(LBP). Similarly, for the tasks of landmarks localization, age and gender estimation and attributes classification, different hand-crafted features were derived. However, with the rise of deep learning, the hand-crafted features were replaced by DCNN features for solving the above mentioned facial analysis tasks. This made it possible to train a single DCNN model for multiple tasks such as face detection, landmarks localization, facial attributes prediction and face recognition.

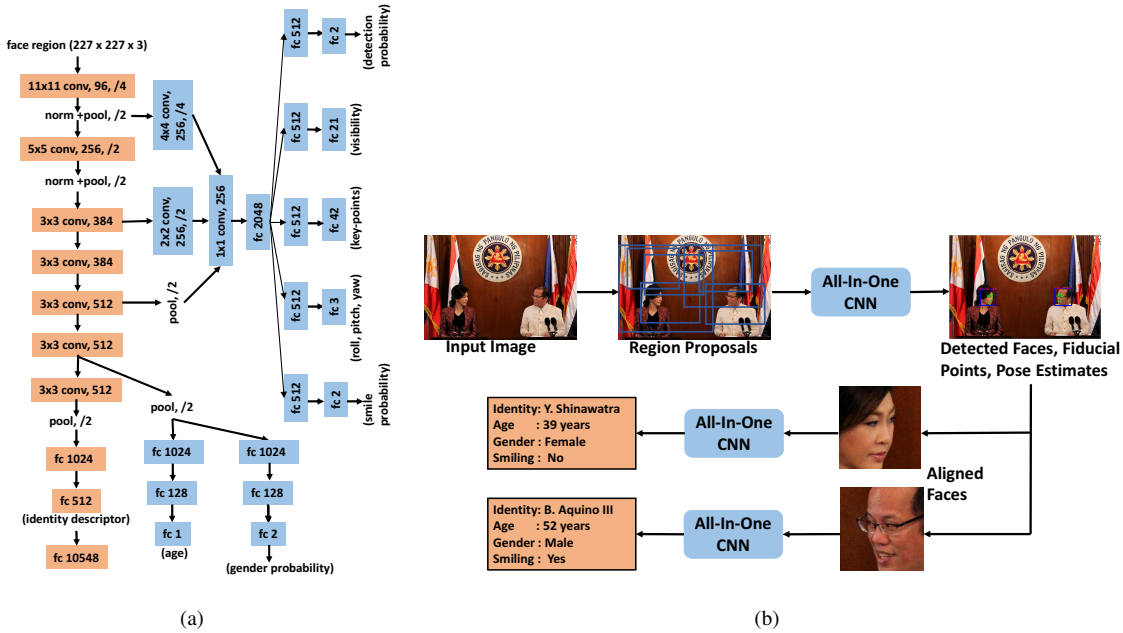
In general, when a human looks at a face in an image, he/she can detect where the face is, determine the gender, rough pose, age, expressions, etc. When machines are designed to perform these tasks, one often constructs independent algorithms for solving each of these tasks. However, one can design a deep network that can simultaneously accomplish all of these tasks by sharing the deep features and exploiting the relationships among these tasks. Goodfellow *et al.* [106] interprets MTL as a regularization method for DCNNs. With the MTL approach, the learned parameters of network are in consensus with all the tasks in hand, which reduces overfitting and converges to a robust solution.

HyperFace [10] and TCDCN [107] are some of the earliest methods to use DCNNs in the MTL framework. HyperFace [10] was proposed for solving the tasks of face detection, landmarks localization, head-pose estimation and gender classification. It fused the intermediate layers of a DCNN such that the tasks could leverage semantically rich features from the deeper layers as well as location-specific features from the lower layers. The synergy developed by the MTL approched improved the performance of individual tasks. Zhang *et al.* [107] proposed the TCDCN algorithm that used MTL-based DCNN for facial landmark detection, along with the tasks of discrete head yaw estimation, gender recognition, smile and glass detection. In their method, the predictions for all theses tasks were pooled from the same feature space. They showed that using auxiliary tasks such as glass detection and smile prediction improved the landmarks localization for specific parts of the face.

Ranjan *et al.* recently proposed All-In-One Face [2], which is a single DCNN model for simultaneous face detection, landmark localization, face recognition, 3D head pose estimation, smile detection, facial age estimation, and gender classification. The All-In-One Face architecture (illustrated in Figure 8(a)) starts with the pre-trained face identification network from Sankaranarayanan *et al.* [73]. The network consists of seven convolutional layers followed by three fully connected layers, and it is used as a backbone network for training the face identification task and for sharing the parameters of its first six convolution layers with other face-related tasks. The central tenet is that a CNN pre-trained on face identification task provides better initialization for a generic face analysis task, since the filters retain discriminative face information.

To leverage multiple datasets with all the annotations for face bounding box, fiducial points, pose, gender, age, smile and identity information, multiple sub-networks are trained with respect to task-related datasets, and share the parameters among them since no single dataset contains annotations for all the facial analysis tasks at the same time. In this way, the shared parameters adapt to the complete set of domains instead of fitting to a task-specific domain. At test time, these sub-networks are fused together into a single All-In-One Face. Table V lists the different datasets used for training the All-In-One Face. Task-specific loss functions are used to train the complete network end-to-end. Some sample outputs for All-In-One Face are shown in Figure 9.

MTL-based DCNNs have also been used for determining multiple facial attributes. Dehghan *et al.* proposed



(a) (b)

Fig. 8. (a) The All-In-One Face network architecture and (b) the overview of the All-In-One Face system for face recognition and facial analysis.

TABLE V

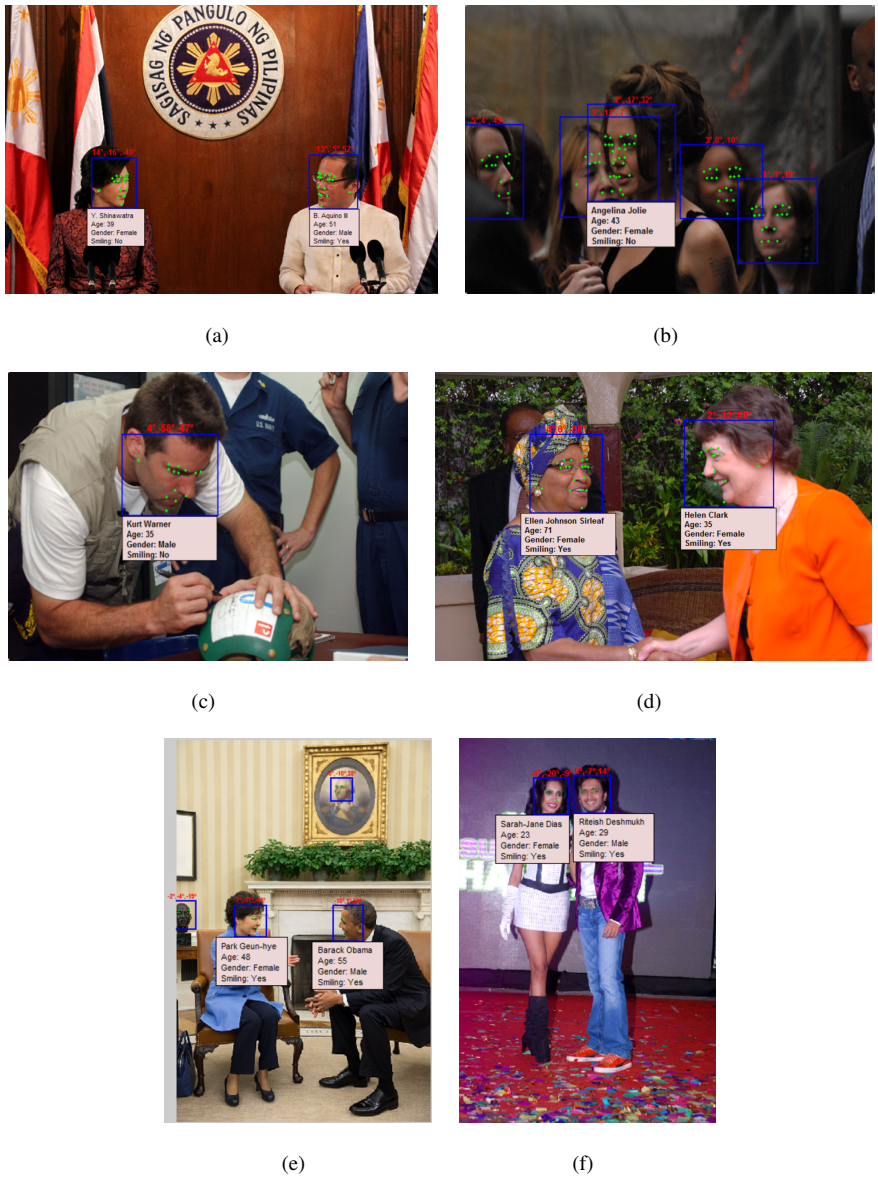
DATASETS USED FOR TRAINING ALL-IN-ONE FACE.

Dataset	Facial Analysis Task	# training samples
CASIA [13]	Identification, Gender	490,356
MORPH [108]	Age, Gender	55,608
IMDB+WIKI [109]	Age, Gender	224,840
Adience [97]	Age	19,370
CelebA [80]	Smile, Gender	182,637
AFLW [110]	Detection, Pose, Fiducials	20,342
Total		**993,153**

DAGER [111] for recognizing age, gender and emotion using DCNN. Similar to All-In-One Face [2], it uses different training sources labeled for different tasks for training the DCNN. He *et al.* [112] jointly learn a deep architecture for face detection and facial attribute analysis. Unlike other MTL-based methods, they use full image as an input to the network instead of a face region. A Faster R-CNN based approach is used to obtain the detections along with the attributes for the faces. Table VI summaries the facial analysis tasks performed by some of the recent MTL-based methods.

VII. OPEN ISSUES

Given sufficient number of annotated data and GPUs, DCNNs have been shown to yield impressive performance improvements. Still many issues remain to be addressed to make the DCNN-based face recognition systems robust and practical. Below, we briefly discuss several design considerations for each component of an automated face identification/verification system, including



(a) (b) (c) (d) (e) (f)

Fig. 9. Sample results of the All-In-One Face [2] for the IJB-A dataset [69] with detected face bounding boxes, fiducial points, identity along with 3D head pose, gender, smile, and facial age estimation. Although the algorithm predicts identity, age, gender and smile attributes for all the faces, we show them only for subjects that are present in the IJB-A dataset for better image clarity.

- **Face detection:** In contrast to generic object detection task, face detection is more challenging due to the wide range of variations in the appearance of faces. The variability is caused mainly by changes in illumination, facial expression, viewpoints, occlusions, etc. Other factors such as blur and low-resolution challenge the face detection task.
- **Fiducial detection:** Most datasets only contain few thousands images. A large scale annotated and unconstrained dataset will make the face alignment system more robust to the challenges, including extreme pose, low illumination, small and blurry face images. Researchers have hypothesized that deeper layers of CNNs can encode more abstract information such as identity, pose, and attributes; however, it has not yet been thoroughly

TABLE VI

LIST OF VARIOUS MTL-BASED FACIAL ANALYSIS ALGORITHMS ALONG WITH THE TYPES OF FACE TASKS THEY CAN PERFORM

Method	Face Detection	Fiducials	Head-Pose	Gender	Age	Expression	Other Attributes	Face Recognition
Zhu *et al.* [55]	✓	✓	✓					
JointCascade [105]	✓	✓						
Zhang *et al.* [32]	✓	✓						
TCDCN [107]		✓	✓	✓		✓	✓	
HyperFace [10]	✓	✓	✓	✓				
He *et al.* [112]	✓			✓	✓	✓	✓	
DAGER [111]				✓	✓	✓		
All-In-One Face [2]	✓	✓	✓	✓	✓	✓		✓

studied which layers exactly correspond to local features for fiducial detection.

- **Face identification/verification:** For face identification/verification, the performance can be improved by learning a discriminative distance measure. However, due to memory constraints limited by graphics cards, how to choose informative pairs or triplets and train the network end-to-end using online methods (*e.g.*, stochastic gradient descent) on large-scale datasets is still an open problem. Another challenging problem to address is to incorporate full motion video processing in deep networks for enabling video-based face anaytics.

VIII. CONCLUSION

In this paper, we presented an overview of recent developments in designing an automatic face recognition system. We also discussed about various multi-task CNN-based approaches for face analytics. The examples shown in the paper demonstrate that subject-independent tasks benefit from multi-task learning, network initialization from face recognition task, and how these tasks also help to improve face recognition performance.

Given sufficient amount of annotated data and GPUs, DCNNs have been shown to yield impressive performance improvements. Still many issues remain to be addressed to make the DCNN-based systems robust and practical, such as reducing reliance on large training data sets, handling data bias and degradation in training data, incorporating domain knowledge, reducing the training time when the network goes deeper and wider, and building the theoretical foundations to understand the characteristics and behaviors of DCNN models. We refer the interested readers to [12] for more details about other challenges for DCNN models.

IX. ACKNOWLEDGMENTS

We acknowledge the help of Anirudh Nanduri, Amit Kumar, Wei-An Lin, Pouya Samangouei, Emily M. Hand, and Ching-Hui Chen. This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the

ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

AUTHOR BIOGRAPHIES

Rajeev Ranjan (B.Tech, IIT Kharagpur, 2012) is a Research Assistant at University of Maryland College Park. His research interests include face detection, face recognition and machine learning. He is a recipient of UMD Outstanding Invention of the Year award, 2015.

Swami Sankaranarayanan (M.S, TU Delft, 2012) is a Ph.D. candidate at University of Maryland College Park. His research interests include face analysis and adversarial machine learning.

Ankan Bansal (B.Tech-M.Tech, IIT Kanpur, 2015) is a PhD student at the University of Maryland, College Park. His research interests include multi-modal learning, action understanding, and face analysis. He was awarded the Clark School of Engineering Distinguished Graduate Fellowship, 2015-2016.

Navaneeth Bodla (M.S, UFL, 2014) is a Research Assistant at University of Maryland College Park. His research interests include face recognition, object detection and adversarial image synthesis.

Jun-Cheng Chen (Ph.D., UMD, 2016) is a postdoctoral research fellow at the University of Maryland Institute for Advanced Computer Studies (UMIACS). His current research interests include computer vision and machine learning with applications to face recognition and facial analysis. He was a recipient of ACM Multimedia best technical full paper award, 2006.

Vishal M. Patel (Ph.D., UMD, 2010) is an A. Walter Tyson Assistant Professor in the Department of Electrical and Computer Engineering (ECE) at Rutgers University. Prior to joining Rutgers University, he was a member of the research faculty at the University of Maryland Institute for Advanced Computer Studies (UMIACS). His research interests are in signal processing, computer vision and machine learning with applications to radar imaging and biometrics. He was a recipient of the ONR YIP in 2016 and the ORAU postdoctoral fellowship in 2010.

Carlos D. Castillo (Ph.D., UMD, 2012) is an assistant research scientist at the University of Maryland Institute for Advanced Computer Studies (UMIACS). His current research interests include stereo matching, multi-view geometry, face detection, alignment and recognition.

Rama Chellappa (Ph.D., Purdue University, 1981) is a Distinguished University Professor, a Minta Martin Professor of Engineering and the Chair of Electrical and Computer Engineering Department at University of Maryland, College Park. His current research interests are face and gait analysis, 3-D modeling from video, image and video-based recognition and exploitation, compressive sensing, and hyper spectral processing. He received the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society.

REFERENCES

[1] W. Y. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2017.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

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

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[9] R. Ranjan, V. M. Patel, and R. Chellappa, "A deep pyramid deformable part model for face detection," in *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*. IEEE, 2015, pp. 1–8.

[10] R. Ranjan, V. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *arXiv preprint arXiv:1603.01249*, 2016.

[11] A. Kumar, R. Ranjan, V. Patel, and R. Chellappa, "Face alignment by local deep descriptor regression," *arXiv preprint arXiv:1601.07950*, 2016.

[12] J. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, C. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa, "Unconstrained still/video-based face verification with deep convolutional neural networks," *International Journal of Computer Vision*, pp. 1–20, 2017.

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

[14] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4873–4882.

[15] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 87–102.

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *British Machine Vision Conference*, 2015.

[18] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533.

[19] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," *arXiv preprint arXiv:1606.03473*, 2016.

[20] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[21] Y. Li, B. Sun, T. Wu, and Y. Wang, "Face detection with end-to-end integration of a convnet and a 3d model," *European Conference on Computer Vision (ECCV)*, 2016.

[22] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 122–138.

[23] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis, "Ssh: Single stage headless face detector," *arXiv preprint arXiv:1708.03979*, 2017.

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

[25] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 643–650.

[26] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *IEEE International Conference on Computer Vision*, 2015, pp. 3676–3684.

[27] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5325–5334.

[28] S. Yang, Y. Xiong, C. C. Loy, and X. Tang, "Face detection through scale-friendly deep convolutional networks," *arXiv preprint arXiv:1706.02863*, 2017.

[29] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S˙3 fd: Single shot scale-invariant face detector," *arXiv preprint arXiv:1708.05237*, 2017.

[30] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," no. UM-CS-2010-009, 2010.

[31] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Fine-grained evaluation on face detection in the wild," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–7.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[33] P. Hu and D. Ramanan, "Finding tiny faces," *arXiv preprint arXiv:1612.04402*, 2016.

[34] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 2015.

[35] A. Bansal, C. D. Castillo, R. Ranjan, and R. Chellappa, "The do's and don'ts for cnn-based face verification," *arXiv preprint arXiv:1705.07426*, 2017.

[36] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, 2017.

[37] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou, "A comprehensive performance evaluation of deformable face tracking" in-the-wild"," *International Journal of Computer Vision*, pp. 1–35, 2016.

[38] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3409–3417.

[39] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1685–1692.

[40] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.

[41] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[42] X. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.

[43] E. Antonakos, J. Alabort-i Medina, and S. Zafeiriou, "Active pictorial structures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5435–5444.

[44] A. Jourabloo and X. Liu, "Pose-invariant 3d face alignment," in *IEEE International Conference on Computer Vision*, 2015, pp. 3694–3702.

[45] A.Jourabloo and X. Liu, "Large-pose face alignment via cnn-based dense 3d model fitting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4188–4196.

[46] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 146–155.

[47] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476–3483.

[48] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks for real-time face alignment." in *European Conference on Computer Vision (ECCV)*, 2014, pp. 1–16.

[49] X. Xiong and F. D. la Torre, "Global supervised descent method," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2664–2673.

[50] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4177–4187.

[51] A. Bulat and G. Tzimiropoulos, "Convolutional aggregation of local evidence for large pose face alignment," 2016.

[52] A. Kumar, A. Alavi, and R. Chellappa, "Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors," *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017.

[53] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[54] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4295–4304.

[55] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2879–2886.

[56] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Advances in face detection and facial image analysis*, 2016, pp. 189–248.

[57] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2518–2525.

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

[59] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10000 classes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1891–1898.

[60] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[61] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," *arXiv preprint arXiv:1412.1265*, 2014.

[62] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *arXiv preprint arXiv:1503.03832*, 2015.

[63] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassne, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajana, R. Nevatia, and G. Medioni, "Face recognition using deep multi-pose representations," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[64] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" *arXiv preprint arXiv:1603.07057*, 2016.

[65] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *arXiv preprint arXiv:1607.05427*, 2016.

[66] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 499–515.

[67] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[68] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[69] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

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

[71] N. Bodla, J. Zheng, H. Xu, J.-C. Chen, C. Castillo, and R. Chellappa, "Deep heterogeneous feature fusion for template-based face recognition," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[72] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1875–1882.

[73] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," *arXiv preprint arXiv:1604.05417*, 2016.

[74] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.

[75] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[76] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Joint face representation adaptation and clustering in videos," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 236–251.

[77] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller, "A deep matrix factorization method for learning attribute representations," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 417–429, 2017.

[78] W.-A. Lin, J.-C. Chen, and R. Chellappa, "A proximity-aware hierarchical clustering of faces," *IEEE Conference on Automatic Face and Gesture Recognition*, 2017.

[79] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large scale face recognition," in *European Conference on Computer Vision*.

[80] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[81] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition*, 2008.

[82] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[83] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013, pp. 1–8.

[84] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[85] A. Bansal, A. Nanduri, C. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks," *arXiv preprint arXiv:1611.01484*, 2016.

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

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

[88] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *arXiv preprint arXiv:1509.00244*, 2015.

[89] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 529–534.

[90] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[91] "National institute of standards and technology (NIST): IARPA Janus benchmark-a performance report," 2016.

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

[93] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017.

[94] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[95] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?" in *Defense and Security*. International Society for Optics and Photonics, 2004, pp. 561–572.

[96] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644.

[97] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[98] P. Samangouei and R. Chellappa, "Convolutional neural networks for attribute-based active authentication on mobile devices," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–8.

[99] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 801–808.

[100] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[101] M. Günther, A. Rozsa, and T. E. Boult, "Affact-alignment free facial attribute classification technique," *arXiv preprint arXiv:1611.06158*, 2016.

[102] P. Samangouei, V. M. Patel, and R. Chellappa, "Attribute-based continuous user authentication on mobile devices," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015, pp. 1–8.

[103] P. Samangouei, V. Patel, and R. Chellappa, "Facial attributes for active authentication on mobile devices," *Image and Vision Computing*, vol. 58, pp. 181–192, 2017.

[104] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[105] Y. W. X. C. D. Chen, S. Ren and J. Sun, "Joint cascade face detection and alignment," in *European Conference on Computer Vision*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014, vol. 8694, pp. 109–122.

[106] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[107] Z. Zhang, P. Luo, C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*, 2014, pp. 94–108.

[108] K. Ricanek and T. Tesafaye, "Morph: a longitudinal image database of normal adult age-progression," in *International Conference on Automatic Face and Gesture Recognition*, April 2006, pp. 341–345.

[109] R. Rothe, R. Timofte, and L. V. Gool, "Dex: Deep expectation of apparent age from a single image," in *IEEE International Conference on Computer Vision Workshop on ChaLearn Looking at People*, 2015, pp. 10–15.

[110] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[111] A. Dehghan, E. G. Ortiz, G. Shu, and S. Z. Masood, "Dager: Deep age, gender and emotion recognition using convolutional neural network," *arXiv preprint arXiv:1702.04280*, 2017.

[112] K. He, Y. Fu, and X. Xue, "A jointly learned deep architecture for facial attribute analysis and face detection in the wild," *arXiv preprint arXiv:1707.08705*, 2017.