Trends in Machine and Human Face Recognition

Bappaditya Mandal, Rosary Yuting Lim, Peilun Dai, Mona Ragab Sayed, Liyuan Li, and Joo Hwee Lim

Abstract Face recognition (FR) is a natural and intuitive way for human beings to identify or verify or at least get familiar and interact with other members of the community. Hence, human beings expect and endeavor to develop similar competency in machine recognition of human faces. Due to the rapid increase in computing power in recent decades and the need to automate the FR tasks for many applications, researchers from diverse areas like cognitive and computer sciences are making efforts in understanding how humans and machines recognize human faces respectively. Its application is innumerable (like access control, surveillance, social interactions, e-commerce, just to name a few). In this chapter we will review two aspects of FR: machine recognition of faces and how human beings recognize human faces. We will also discuss the recent benchmark studies, their protocols and databases for FR and psychophysical studies of FR abilities of human beings.

1 Introduction

Among many biometrics, such as finger print, palm print, ear, iris, gait, etc., face is considered to be most user-friendly and intuitive as the authentication can be performed at a distance, even without the knowledge or cooperation of the subject. The main difficulties that face recognition (FR) algorithms have to deal with are two types of variations: intrinsic factors (independent of viewing conditions) such as age and facial expressions and extrinsic factors (dependent on viewing conditions) such as pose and illumination. Large amount of work has been done over the last three decades to address these issues. Starting from the pioneering work of Eigenfaces by Pentland et al. [1] to the latest results of DeepFace by Wolf et al. [2] and DeepID by Wang et al. [3], researchers are able to reduce the recognition error rate (%) from two digits to near perfection [4].

B. Mandal (✉) • R.Y. Lim • P. Dai • M.R. Sayed • L. Li • J.H. Lim
Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, #21-01, Connexis (South Tower), Singapore 138632, Singapore
e-mail: bmandal@i2r.a-star.edu.sg

M. Kawulok et al. (eds.), *Advances in Face Detection and Facial Image Analysis*,
DOI 10.1007/978-3-319-25958-1_7

Although high recognition rates are reported in the academic papers, there are large gaps between the reported performance in constrained framework and their performances in the large scale unconstrained environment. In this study, we will discuss the recent face recognition vendor test (FRVT 2013) [5] conducted by the National Institute of Standards and Technology (NIST) as a third party independent evaluator of FR algorithms. We will discuss these gaps, the difference in old and new benchmark protocols and results reported in the recent benchmark study of large-scale unconstrained FR by Stan Z. Li et al. [6] on the well-known 'labeled faces in the wild' (LFW) database in 2014. Unlike numerous previous studies on LFW, this large scale unconstrained FR algorithm evaluation with new protocol using entire LFW database reveals that only 41.66 % correct verification rates (CVR) can be obtained at 0.1 % false acceptance rates (FAR) and 18.07 % open-set face identification (FI) rates at rank 1 and 1 % FAR. As these numbers show that FR problem is still largely unsolved, we will devote more attention and efforts in reviewing new invariant feature representations and learning algorithms that can advance the algorithm development for FR.

In addition to machine recognition of faces, we will review how human beings perceive human faces for recognition. We respond to faces differently from other classes of objects. Interaction involves a certain level of social cognition that needs to be adapted for each situation. Research on human face processing is now moving away from the use of static face stimuli and delving into dynamic faces to simulate a more realistic context for FR and processing. This endeavor gave rise to the formulation of two popular hypotheses by O'Toole et al. [7], in an attempt to explain the benefits that dynamic faces impart on human recognition of faces: the supplemental information hypothesis and representation enhancement hypothesis. However, both hypotheses are unable to explain how humans are able to learn and recognize a face with much fewer templates than machines. What are the possible strategies that could optimize learning of novel faces, even under challenging conditions? In this chapter, we analyze the findings for human psychophysics experiments that investigated human performance in FR across varying conditions of illumination, expression, viewing perspectives, and time lapses in age. We suggest possible FR strategies utilized by humans that could be incorporated into machines to pave the way for next-generation recognition systems.

In Sect. 2, we will discuss briefly the challenges involved in FR and its general pre-processing and normalization steps. In this section, we will also study the dynamics of FR in unconstrained environment involving emerging techniques. We will do a review on FRVT 2013 and some of the emerging databases, their findings, protocols and summary in Sect. 3. Motivated by the limitations of machine recognition of faces as discussed before, we also do a comprehensive review for dynamics in human recognition of faces in Sect. 4. It would help us to understand how we human beings solve these problems and challenges posed by machines. We also share the experimental results of human performances in dynamic FR. Finally in Sect. 5, we conclude and discuss the future trends.

2 Machine Face Recognition: Its Existing Challenges and Emerging Methods

Human face recognition plays an important role in our daily life. We utilize our visual memory to recognize an individual [8] or at least able to recall seen / unseen (familiar or unfamiliar) individual faces [9]. For humans, FR is the most natural and common way to identify and/or verify individuals. It is so intuitive and non-intrusive (without user intervention) that we aim to replicate this capability into machines. Even after four decades of intensive research in machine FR, the problem is still far from solved for large scale unconstraint FR. So what are the existing challenges that extirpate us from achieving human like high recognition rates?

2.1 Challenges for Face Recognition

For unconstrained FR, the challenging factors are: *Illumination variation, Pose and viewpoint variation, Expression variation, Aging, Scale variation, Occlusions and Motion blur*. One or in combination with others, have caused tremendous challenging problems for large scale unconstrained FR. Below we discuss each of these problems briefly.

2.1.1 Illumination Variation

A person's face appears quite different at different times throughout the course of a video capturing when it passes through underneath lights or some strong lights in certain directions. Illumination also results in self shadowing making the problem even harder. Some samples images of a person with different illumination conditions from YaleB database [10] are shown in Fig. 1. A large amount of research work has been devoted to study and alleviate this problem [11–13], however, all of them studied the problem of illumination under constrained (studio settings) environment. Very few studies are performed on real-life unconstrained illuminating conditions [14].

2.1.2 Pose and Viewpoint Variation

In natural settings, either the subject or/and viewer are moving. Capturing facial images from a stationary or moving (wearable devices like Google Glass) camera, the moving faces can lead to shots from a variety of angles causing the correspondences between pixel locations and points on the face to differ from image to image. Since human face is a 3D structure, using only 2D images to reconstruct unknown poses can become an ill-posed problem. Camera capturing human face

Fig. 1 Appearance of a person under different illuminating conditions from YaleB database. 'A+035E+15' implies that the light source direction with respect to the camera axis is at 35° azimuth ('A+035') and 15° elevation ('E+15'). (Note that a positive azimuth implies that the light source was to the *right* of the person while negative means it was to the *left*. Positive elevation implies above the horizon, while negative implies below the horizon)

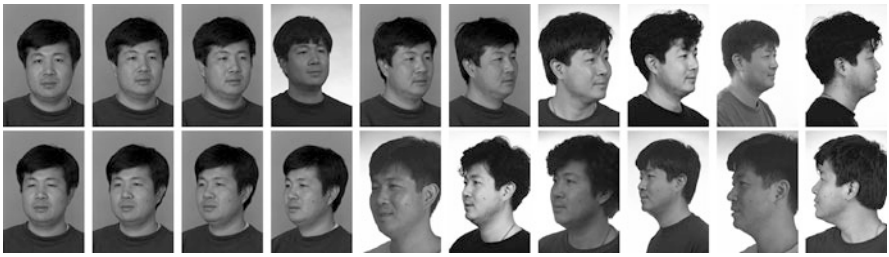

Fig. 2 Sample images from the FERET database [15] for one person with varying degree of poses

images results in in-plane or out-of-plane rotations as shown in Fig. 2. The former is a pure 2D problem and can be solved much more easily, like placing the eyes on the same horizontal axis [16]. However, the latter is very challenging and is also known as in-depth rotations. When part of a face is invisible in an image due to rotation in-depth, the facial texture is recovered from the visible side of the face using the bilateral symmetry of faces. Human face is limited to three degrees of freedom in pose, which can be characterized by pitch, roll and yaw angles. Extracting accurate face pose information in terms of these angles has always been a very challenging problem in FR literature [17, 18].

2.1.3 Expression Variation

Although all faces share the same configuration of two eyes, a nose and a mouth, forehead and cheek regions, signification in-depth deformations occur because of our expressions. Some sample images from AR database [19] are shown in Fig. 3. They pose serious problems to FR performance [20].

Fig. 3 Sample images from the AR database for one person with 14 different expressions

Fig. 4 Sample images of one person in different ages

2.1.4 Aging

Human face changes considerably along with aging, it gets effected in different forms at different ages. During one's younger years the cranium's shape of the face gets more effected whereas they are more effected in terms of wrinkles and other skin artifacts during one's older age. Human face also undergo growth related changes and changes arising from environmental effects that are manifested in the form of textural, color and shape variations. Some sample images of a person in different ages (from [21] database) are shown in Fig. 4. Extracting features that are invariant to large variations in ages for FR is a very challenging problem. Moreover, collecting and archiving face images across different ages in different years (decades) itself is non-trivial effort.

2.1.5 Scale Variation

Because of moving cameras and/or moving persons, face images are captured at different scales resulting in different resolutions. Some samples images captured at 1, 2, 3 and 4 m with no zooming condition using Google Glass are shown in Fig. 5. The original image resolution in Google Glass is set to 360×640 and the cropped images shown in Fig. 5 are of 150×140 dimensions each. Existing research shows that a high resolution 2D face image is better for FR than one 3D face image [22].

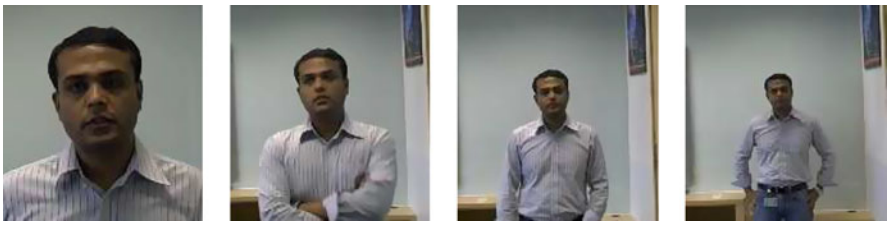

Fig. 5 Sample images of one person captured at 1, 2, 3 and 4 m (*left to right*) with no zooming condition using Google Glass. The face sizes are 90×90, 50×50, 36×36 and 21×21 respectively

Fig. 6 Sample images of three persons, one in normal and two partially occluded conditions (opaque glass and scarp)

Fig. 7 Sample images of three persons with motion blur captured using Google Glass. In all the cases face and eyes detections were successful [23]

2.1.6 Occlusions

Objects in the scene can block a face resulting in reducing the visible area of the face. Common cases like wearing opaque glasses can cause severe occlusions to the eyes areas. Due to occlusions the amount of face information captured is reduced, which makes the FR problem more difficult (Fig. 6).

2.1.7 Motion Blur

Either a moving face or/and a moving camera can cause motion blur. Also when the camera exposure time is set too long or the head moves rapidly, motion blur can occur. Distinctive characteristics of a face are lost when they are blur resulting in poor FR performance, such as in wearable devices [23]. Some samples images are shown in Fig. 7.

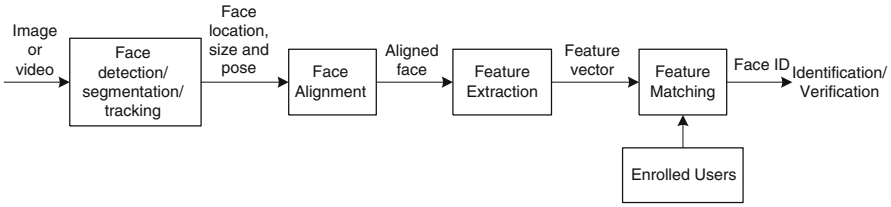

Fig. 8 A general face recognition system

2.2 Pre-processing and Normalization

In the general FR framework as shown in Fig. 8, numerous researchers perform detection of face and its features (like eyes) [16, 24]. Detected face and facial features are used for face alignment. Generally, eyes are placed on the same horizontal axis and at fixed distance (pixels) apart. A face mask is then applied to mask out the non-face portions (like the background) arising above the shoulder and below the chin. It also helps to remove the hair region which has high variations. This whole process is called pre-processing or normalization step. If facial features (like eyes) are detected wrongly then the subsequent processes may fail or the system will achieve very low recognition accuracy. The dependency between the detection precision and recognition accuracy has been studied in [25] by Kawulok et al. In recent years, face and its features detection has been improved to a very large extent. However, for unconstrained scenarios they are still challenging [26, 27]. Low-dimensional features are extracted from high-dimensional objects like face images and stored into the database. When new images (of enrolled users or imposters) are captured, they also undergo similar processes and matching is done with the features stored in the database. Finally, a match ID or non-match (unknown) outcome is given as output.

2.3 Trends in Unconstrained Face Recognition: Promising Directions

Over the past three decades researchers from diverse fields are making efforts in improving the FR algorithms. We have tried to summarize the popular or distinct algorithms that are developed over these years in Fig. 9. It is beyond the scope of this chapter to discuss each of these approaches. For details of the methodologies belonging to holistic, component and hybrid based approaches, the readers are advised to refer the FR survey paper by Zhao et al. [28]. For methodologies grouping based on three levels of taxonomy of facial features, the readers can refer to the paper by Klare et al. [29]. A recent 2014 survey on single and multimodal FR can be found in [30]. There are also a few papers that review FR across pose variations [17, 18], illumination variations [31, 32], aging [33] and forensic applications [34].

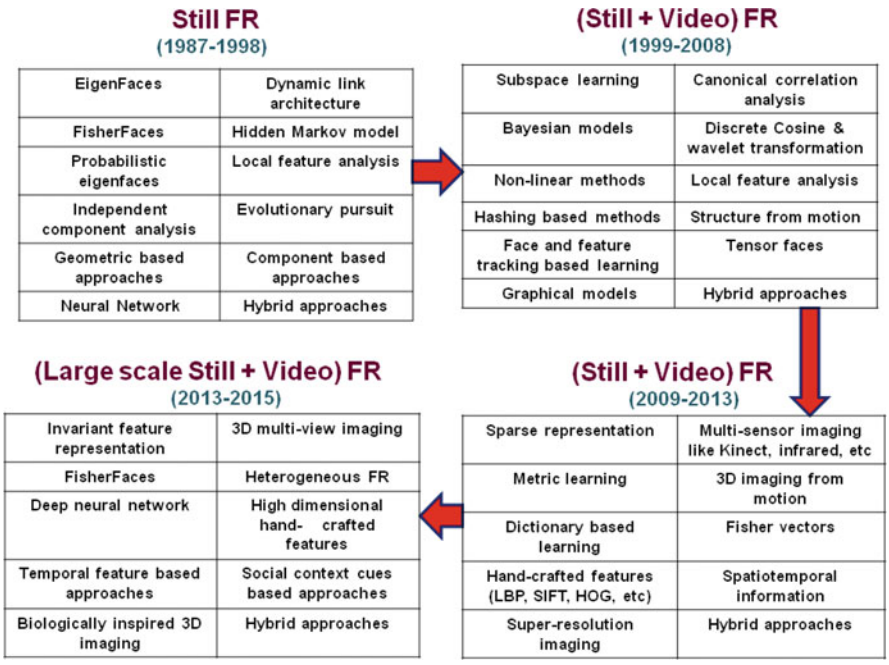

Fig. 9 Evolution of face recognition algorithms

The recent interest in FR is motivated from a few promising directions, which are (1) approaches that use the biologically motivated theory of invariance identity-preserving transformations, (2) video-based FR and (3) deep-learning based convolutional neural-network framework. Below we discuss these promising directions and a few recent successful examples.

2.3.1 Methods Using Invariance Identity-Preserving Transformations

It is evident from the recent literature reviews on unconstrained FR that in order to develop the next generation FR algorithm that can perform better FR as compared to humans and even surpass human performance, we would need more challenging databases as compared to the past. Leibo et al. [21] has tried to come up with a unconstrained FR database which is much more challenging as compared to the previously well studied labeled faces in the wild (LFW) [35] and YouTube face image (YTF) databases [14]. They named this database as subtasks of unconstrained face recognition (SUFR) [21]. Their idea is to isolate faces with specific transformation or a set of transformations for different subtasks to suppress the common computational problem of FR which is transformation invariance to various translations, illuminations, rotations and scalings. Leibo et al. [21] produced six artificial face image datasets using 3D graphics based on this concept, where

each of them contains face images created using a set of transformations with various cluttered/homogenous backgrounds. Although they proposed a good idea to handle the unconstrained conditions resulted from various transformations but they are still using affine transformation. They have not included the difficult variations and deformations like face expression, along with pose and aging. So, this approach is still incomplete and cannot be used in most of the real-world conditions.

Motivated by the recent theory of transformation invariance [36], Liao et al. [37] used the SUFR database for face verification (FV) following the same idea of finding the invariance features using various transformations. A signature or invariant representation for each image is computed with respect to a group of transformations. As the inner product of the image and transformed template is the same as the inner product of the template with the same transformed image, their empirical distribution function of the inner products can be used as signature for each image. Although the authors reported a good performance of this model but it may assume some restrictions where the transformation is non-affine. For example, the authors stated that this model may work in case of non-affine transformation when it is restricted to certain nice class such as the 2D transformation mapping of the face image to its frontal view is similar to transformation of another face within the same scene. Also, as this model depends on the distribution of the inner product of an image and transformed template, it means that the transformation has to be known or can be measured in advance which is not applicable in many practical cases. Finally, it does not consider the background variations in all circumstances.

A vast number of FR research follow the algorithmic flow of *face detection* → *normalization* → *face recognition* [16, 28, 38, 39]. However, the recent theory of invariant recognition by feedforward hierarchical networks [40], like HMAX [41, 42], and other convolutional networks [43], or possibly the ventral stream, implies an alternative approach to unconstrained FR. The main idea is to remove traditional FR pipeline techniques such as face cropping, alignment and normalization and use the whole image (possibly with a face in it) for recognition. This is a biologically motivated way for performing FR as we human beings do not use normalization explicitly while recognizing a face [44].

Liao et al. [44] used a three-levels HW-Module architecture (in honor of Hubel and Wiesel's original proposal for the connectivity of V1 simple and complex cells [45]) to obtain the face signature (identity) of an individual. At level one, face is detected, nearly cropped and low level features are extracted at different positions and scales for each image in the training set. These features are stored in vectors as training templates. Then they compute dense overlapping set of windows for each test image, convolved them with training templates and applied max pooling to get new templates. Finally, matching process is done at the third layer by obtaining the dot product between these templates and the training templates and scores are computed for each test image. In that work, Liao et al. [44] tried to reduce the complexity of these processes by hashing and rank approximation using principal component analysis (PCA). They applied this model on different unconstrained databases like LFW, SUFR-W (SUFR-in the wild), LFW-J (LFW-jittered) to get state-of-the-art FR accuracy rate of 87.55 %. This accuracy is near to other popular

FR approaches that use cropping, alignment and normalization of the testing set [46]. So, they proved experimentally that their biologically plausible hierarchical model can effectively replace face detection, alignment and normalization pipelines [44], however, these techniques are of limited use with non-affine transformations.

2.3.2 Video Based Face Recognition

In the machine face recognition literature, majority of the research has focused on improving the ability of FR using static (still) face images. As pointed out in [47], this is primary because of factors such as (1) the need to constrain FR problem, so that the researchers can focus on specific type(s) of FR problem (one in combination with other, such as illumination and/or expression in AR database [19]) and assume all other factors as more or less constant, (2) computational or hardware constraints for both acquiring, processing and storing large amount of face images, (3) the large amount of legacy still face images (e.g. ID cards, mug shots) and (4) its limited availability (or sharing such as in social networks [48, 49]). Today, many of these constraints are no longer valid. A large number of researchers are working on computational, biological and cognitive aspects of FR [50–52], tackling the problem well and coming up with new model, theory and challenging unconstrained databases [53]. FR using still images has witnessed an exponential decrease in error rates [5]. Hardware devices (like digital cameras) for acquiring or capturing images are becoming less and less expensive. Availability of distributed and parallel computing has helped in processing a very large number of images. Lastly, people are very active in sharing images/videos across multiple domains, internet and channels (like social networks), hence they are more readily available as compared to the past [48, 49].

As described earlier, compared to legacy static (still) images, videos help in enhancing FR process as additional information can come from motion and other aspects such as multiple faces of different poses, expressions and illuminations. Firstly, there are techniques based on feature extraction from video input, such as [54]. These features may represent the relation between facial features or the invariant structural features that do not change under different conditions such as skin-model based and color-based approaches. Also global features are useful, such as shape of the face, skin and size or detail features of the internal face components (like eyes and nose). Secondly, there are methods based on probability density function. They deal with face images as random variables of certain probability where the similarity between images can be measured by similarity of their corresponding probability density function. Thirdly, some techniques use the dynamic variance of faces in images to enhance the face detection and identification by integrating features extracted from sequence of images like motion information.

Rowden et al. [47] proposed two techniques to fuse information from image sequences in unconstrained conditions using YouTube faces database [14]. Their multi-frame fusion deals with video as a group of single still frames. Each frame in the query video is matched to the corresponding video in the database and similarity

score is computed and measured as a part of the verification process. Scores from all images are then combined by averaging, max, min and median rules. Fusion can also be done to combine matchers score before or after the multi-frame fusion using the similar rules. The former is called the multi-matcher multi-frame (MMMF), while the latter is called multi-frame multi-matcher (MFMM) technique. These techniques are tested using three commercial off-the-shelf algorithms. According to their results, the accuracy of identification using frames fusion is better than still images which means that videos are better than still images to recognize faces in unconstrained conditions. Also, fusion of more than one matchers achieves better performance. Although good performances are achieved by these techniques but matching each frame in query video with all frames in the database videos may take large computation time. Also, as the final decision is a result of combining more than one matcher scores, it may lead to failure if some of the matchers scores are very low. This could happen especially in unconstrained conditions such as low resolution and occluded face images.

Li et al. [55] proposed a technique to decrease the complexity of identification on large-scale databases by representing each subject in all relevant videos by one Eigen-PEP (probabilistic elastic part) representation with invariant length over different YouTube face videos. This representation can be used later in the matching process to make identification using joint Bayesian classifier. This approach achieves high performance identification of 85.04 % on YTF dataset and verification rate of 88.97 % on LFW dataset.

Other researchers like Chen et al. [56] tried to exploit the temporal information between video frames using joint sparse representation. They divided the database into various partitions, each partition has images of the same pose and illumination for the same face from the same video. Each video is represented by many partitions which is learned under strict sparsity to find the best representation of each face in each of the partitions. Same methodology is applied for the test images, used in the later matching step. Using this technique, the best identification rate 98.04 % is achieved on UMD dataset [57], where each subject has at least six sequences of images. This technique takes into account the illumination and pose conditions but have not exploited all the unconstrained conditions.

2.3.3 Deep-Learning Framework for Face Recognition

Recently, an emerging class of FR algorithms using large number of diverse yet labeled face images and deep neural nets (DNN) have shown promising recognition performance in unconstrained environment. The generalization capability of many machine learning tools like support vector machines (SVM), PCA, linear discriminant analysis (LDA), Bayesian interpersonal classifier tend to get saturated quickly as the volume of the training increases [58–61]. DNNs have shown to perform significantly better as compared to traditional machine learning algorithms [2] when trained with large number (millions) of diverse images, for example, images appearing in Facebook [49] at different times (and not similar appearing

faces in videos). However, DNNs requires large amount of training data without which the network fails to learn and deliver impressive recognition performance. Moreover, training such massive data requires huge computational resources, like thousands of CPU cores and/or GPUs. Zhu et al. [62] trained DNNs to transform faces from different poses and illumination to frontal faces and normal illumination. They used features from the last hidden layer and transformed the faces for FR. Sun et al. [63] used multiple DNNs to learn high level face similarity features and used restricted Boltzmann machine for FV. They extracted features from a pair of face images instead from a single face.

DeepFace developed by Taigman et al. [2] has become very popular among the FR society. Primarily they have two good contributions in their work. Firstly, a 3D alignment process, where they used 3D modeling of the face based on fiducial points, that is used to warp a detected aligned 2D facial crop to a 3D frontal mode. They extracted fiducial points by using a support vector regressor trained to predict point configurations from local binary pattern (LBP) histograms based image descriptors [64]. For the alignment of faces with out-of-plane rotations, Taigman et al. used a generic 3D shape model and registered a 3D affine camera. Using these they transformed the frontal face plane of the 2D aligned crop to the image plane of the 3D shape [2].

Secondly, Taigman et al. [2] developed an efficient DNN architecture using 4 million images from 4000 persons. Face detection and localization are performed by extracting 67 fiducial points on each of the face images. Then, triangulation and frontalization are done to 3 RGB layers which are feed into 32 filters (convolutional layer 1: C1) as shown in Fig. 10. The output of this step includes 32 feature maps. M2 layer is a max pooling to get the maximum of these maps over 3×3 spatial neighborhood. Convolutional layer 3 (C3) contains filters which extract the low level features. So, C1, M2, C3 are responsible for features extraction. Three layers (L4, L5 and L6) are used to apply filter bank where every location in feature map learn a set of filters. Finally, the last two layers are connected to get the correlation between the features extracted. This DNN involves more than 120 million parameters using several locally connected layers without sharing weight, unlike the standard convolutional layers. DeepFace when applied to LFW and YTF databases achieves and an impressive accuracy rate of 97.35 % and 92.5 % respectively.

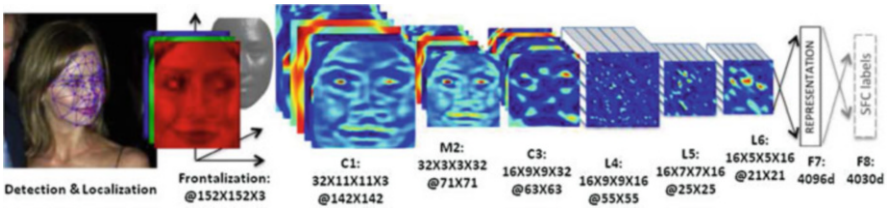

Fig. 10 DeepFace learning framework, from Taigman et al. [2] (Best viewed in *color*)

Another notable deep convolutional network (ConvNets) architecture called DeepID is developed by Sun et al. [3]. It contains four convolutional layers (with max-pooling) to extract features hierarchically. DeepID features are taken from the last hidden layer neuron activations of the ConvNets, followed by the softmax output layer indicating identity classes. Weakly aligned face image patches are used as inputs to each of the ConvNet, which extracts local low-level features. Number of extracted features gets reduced along the feature extraction hierarchy until the last hidden layer (DeepID layer) is reached. In this DeepID layer low dimensional predictive features are formed, which can predict an impressive 10,000 identity classes [3]. They have pioneered CelebFaces and CelebFaces+ face databases. The latter being a superset of the former contains 202,599 face images of 10,177 celebrities from the Internet. People in LFW and CelebFaces+ are mutually exclusive. Using their proprietary (not publicly available) databases and highly compact 160-dimensional DeepID features, they could achieve 97.45 % face verification accuracy on LFW, using weakly aligned face images [3].

Although, there are great and promising performance enhancement in these works, they still need to deal with very large scale evaluation on unconstrained FR (described in the Sect. 3.2.3: old and new protocol on LFW) in order to get good results. In the next section we review some of the benchmark competitions and evaluations done by independent organization and large research organizations.

3 Evaluation and Benchmark Competitions

The development of FR technology has started in 1993 and over of the period of time it has evolved to a very large extend including its applications from large scale nationwide deployment to ubiquitous wearable device computing. We have tried to summarize the entire evolution of FR benchmarks, competitions and algorithm evaluations in Fig. 11. The first FR technology test [65] took place in 1996 and this has lead to multiple FR vendor test conducted in 2000, 2002, 2006 and 2013. Links to all these FR vendor tests can be found in [66]. In between there are other competitions that took place, which include face recognition grand challenge (FRGC) 2005 [67], multiple biometric grand challenge (MBGC) 2009 [68], face and ocular challenge (FOCS) 2009 [69], good, bad and the ugly face challenge problem (GBU) 2009 [70] and multiple biometric evaluation (MBE) 2010 [71]. Furthermore, FR technology evaluation has been extended to mobile devices/environment like MOBIO in 2013 [72]. Generally, each of these competitions and evaluations takes place over 1–3 years time. Recently, due to the increasing popularity of social network and inexpensive "point-and-shoot" camera technology, people would just want to take pictures or videos, upload and recognize their friends, family and their acquaintances more-or-less automatically. This has spurred the point and shoot FR challenge (PaSC) in 2015 [73]. Going into details of each of them is beyond the scope of this chapter. In this section, we review some of the benchmark

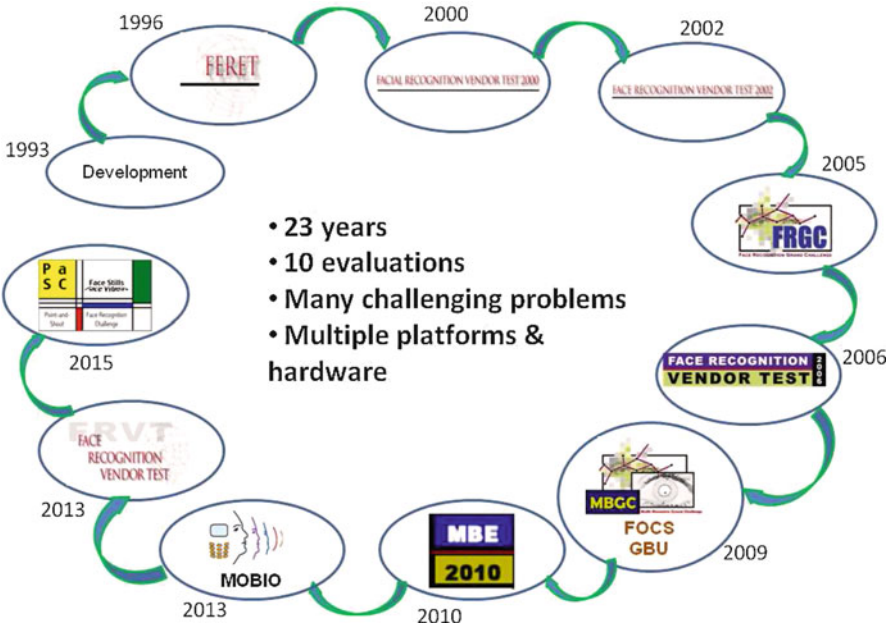

Fig. 11 Face recognition benchmarks, competitions and algorithm evaluations. (Best viewed in *color*)

evaluations and competitions that took place over the last few years, their protocols and summary of the evaluations. Later, we also discuss the emerging databases resulting from these competitions and evaluations.

3.1 FRVT 2013 Findings and Conclusions

The Face Recognition Vendor Test (FRVT) is a series of public evaluations for FR systems built by leading FR technology vendors. FRVT has been organized by the National Institute of Standards and Technology (NIST) in 2000, 2002, 2006 and 2013. FRVT succeeded the previous FERET evaluations held in 1994, 1995 and 1996 [15]. The latest one was FRVT 2013 [5], which started in middle 2012 and lasted until the mid of 2014. FRVT uses a large database to test both the accuracy and computational efficiency of various FR algorithms. The database consists of three parts. The first part is the law enforcement images (LEO) mugshot faces, which comprises about 86 % of the LEO database. The remaining 14 % images of the database were recorded by a webcam, which is referred to as LEO webcam. In addition, a smaller set of visa images consisting of well controlled frontal photographs of adults and children is also used. Besides the three types of face images, some sketch images based on the FERET dataset were also collected

to support research in face sketch synthesis and recognition [74, 75]. For the competition, there are five tracks for the participants to participate:

1. Class A: Compare one-to-one verification (determine if two samples originate from the same person or not) accuracy.
2. Class B: Compare one-to-one verification accuracy but with an enrollment database present. This track was discontinued after the 2010 evaluation. Accuracy gains over class A are available.
3. Class C: Compare one-to-many identification (search to determine either that the person is not enrolled, or to determine the identity of the person). The FRVT test only evaluates on "open-set" identification algorithms because real-world applications are usually "open-set". Here, the "open-set" refers to the situation where a test face image might not be enrolled. The various partitions with numbers of enrolled individuals are 20,000, 160,000, 640,000 and 1,600,000.
4. Class D: Compare accuracy of determining the sex or age of a person in one or more input images. This separate class D track tests on determining whether the face in an image is frontal or non-frontal.
5. Class F: Find effectiveness of the algorithms that take one or more non-frontal images of a person as inputs and outputs one or more frontal images of the same person.
6. Class V: Find effectiveness of the algorithms that execute one-to-many identification of persons with frames extracted from surveillance video sequences.

The error measures used in FR evaluations such as for Class C are usually false alarms (search data from a person who has never been seen before is incorrectly associated with one or more enrollees' data) and misses (a search of an enrolled person's biometric does not return the correct identity).

From the results of algorithms submitted to the FRVT 2013 [5] for evaluation from various commercial vendors (NEC, Cognitec, etc.), the following points are observed:

(1) The age of the subjects strongly affects the identification accuracy. For all the algorithms evaluated, the older the person, the easier they are to be recognized. For children, both false alarm rate and miss rate are higher than other age groups. And infants are very difficult to identify.
(2) Sketch images are also used to match face photographs. For the most accurate algorithms, the rates of face not being among the top 50 candidates are quite high with 73.3 % for 3M/Cogent, 73.8 % for NEC, 78.5 % for Toshiba, 80.3 % for Morpho and 81.5 % for Neurotechnology.
(3) The image quality improvement is the largest contributor to the increase in recognition accuracy. The results show that there is a fourfold reduction in miss rate using high-quality mugshots vs. low-quality webcam images.
(4) The 2010 NIST FR evaluation showed that retention and use of all historical images increase accuracy considerably [76].

The FRVT 2013 provides independent evaluations of commercially available FR systems. These evaluations are aimed at helping the U.S. government agencies best

evaluate and determine the scenarios where these technologies can be deployed. It also helps the FR research community to identify the limitations of current FR technologies and future research directions for improvement. As for the limitation of the dataset provided in FRVT 2013, neither the mughots nor the visa images have ideal properties. The mugshot images have too much pose variation while the visa images are degraded by the acquisition process and the JPEG compression.

3.2 Emerging Databases

3.2.1 The Good Bad and the Ugly (GBU) Datasets

In the past four decades, performance of FR on frontal still faces in controlled environment has improved significantly and achieved near perfect performance. However, frontal faces taken with uncontrolled environment (illumination) and expression remain challenging. As part of the Face and Ocular Challenge Series (FOCS) [69, 76], the Good, the Bad, and the Ugly (GBU) dataset tries to encourage algorithms that work well on matching "hard" face pairs but not at the expense of the performance on "easy" face pairs [70]. The GBU dataset consists of three partitions of frontal still face images, "Good", "Bad" and "Ugly". The three partitions were of different "difficulty" levels with the "Good" being the easiest partition, "Bad" being the average difficult partition and "Ugly" being the most challenging partition based on the analysis of results of the FRVT 2006 challenge [77, 78]. Some sample images are shown in Fig. 12.

Each of the three partitions has two sets, the target set and the query set. Each of the target and query sets in the three partitions contains 1085 images for 437 distinct people, 117 people with one image, 122 people with two images, 68 people with three images and 130 people with four images. The fusion algorithm based on the fusion of the top three performers of FRVT 2006 [78] were used to evaluate the similarity of face pairs and construct the three partitions. For FR, many factors contribute to the recognition performance with the big four factors being subject aging, pose variation, illumination and expression. The GBU dataset controls for subject aging, pose variation and the major factors that affect recognition are illumination and expression, as shown in Fig. 12. In order to avoid over-fitting on the data, the protocol of GBU does not allow training on images of subjects in the GBU dataset. A baseline algorithm, Local Region PCA (LRPCA) [70] is presented and evaluated to illustrate the training and evaluation protocol and provide a baseline performance for comparison.

Besides the original goal of stimulating research on "hard" FR problems, the GBU dataset can also be used to study other factors that could contribute to improving FR performance such as in [79]. In [79], the GBU dataset has been used to study the demographic effects on estimates of automatic FR performance. Based on their findings, the measures of FR performance rely both on the distribution of faces of matched identity as well as mismatched identities. They showed that the

Fig. 12 Sample face images of 1 person from the GBU dataset, from Sinha et al. [4, 69]. The Good pair is referred to "Good", challenging pair as "Bad" and very challenging pair as "Ugly" (Best viewed in *color*)

demographic diversity differences in the non-matching distribution can radically change the estimates of FR algorithm performance. Thus, it poses a new challenge to find a method for tuning algorithm performance to the changing demographic environments where these FR systems will be used reliably.

3.2.2 FR in Mobile Environment

The MOBIO database provides the FR community a bi-modal (audio and video) dataset recorded in a less controlled environment by mobile phones. It also comes with an evaluation protocol together with a baseline algorithm to compare different algorithms developed by the participants of the FR competition in mobile environments hosted at the 2013 International Conference on Biometrics [80, 81]. The goal of the dataset is to stimulate research in the field of multi-modal recognition in a mobile environment. For first-person-view (FPV) or egocentric views face images, Mandal et al. [27] reported a database comprising of face images captured using wearable devices like Google Glass and head mounted web cameras.

Fig. 13 Sample face images of two persons from the MOBIO dataset. There are large variations in pose, illumination, makeups and hair style. (Best viewed in *color*)

MOBIO database was mostly collected by mobile phones with subjects speaking to a handhold mobile phone by answering a set of predefined questions as described in [80]. In total, the dataset consists of 61 h of audio-visual data recorded over a period of one and a half years. The participants consist of 100 males and 52 females, each of whom has 192 unique audio-video samples. For each participant, two phases were recorded, each of which contains six sessions of recordings, and the sessions are separated by several weeks. Some sample images are shown in Fig. 13.

Since the videos are recorded by mobile phones, the dataset has created the following challenges:

- The pose and illumination conditions vary across different samples,
- The quality of the speech recorded varies and
- The environments in which the videos are recorded vary in terms of illumination, background and acoustics.

For evaluation, the dataset is split into three non-overlapping partitions for training, development and evaluation. The training set is used to train the models, e.g. the project matrices for PCA. These images can be used as negative examples in a classification system for some systems. They can also be used for score normalization in training and testing. The development set is used to tune some meta-parameters of the models, e.g. the dimension of the PCA projection matrices. The evaluation set is used to test the models with data that haven's been seen in the training and tuning steps. As the goal of the dataset is to evaluate FR rather than face detection, the eye locations of some selected frames in each video are hand-labeled and provided to the participants.

The organizer provided a baseline algorithm for both speaker recognition and FR, and an algorithm based on fusion of the two modalities (video and audio) is also provided [81]. The baseline algorithm can process 15 frames per second and is suitable for running on mobile devices. For the competition, eight institutions participated and most of the algorithms submitted relied on one or more features of: local binary patterns, Gabor wavelet responses (especially Gabor phases) and color information. With score fusion, the University of Ljubljana and Alpineon Ltd. (UNILJ-ALP) performed best, achieving an equal error rate (ERR) of 2.751 % and 1.707 % on females and males respectively. Among those without fusion algorithms,

the University of Campinas and Harvard University (UC-HU) team achieved the best performance of 4.709 % and 3.492 % on females and males, respectively, without relying on handcrafted features, but learned features with a convolutional neural network [82] instead.

The contribution of the dataset is threefold: first, it provides a challenging FR dataset with uncontrolled face videos; second, the dataset provides both audio and video for fusing the two modalities to improve the identity authentication performance; third, the whole dataset is recorded in mobile phones, and the evaluation requires a trade-off between performance and hardware requirement, which encourages algorithms designed for mobile devices.

The main drawback of the dataset is that in the FR evaluation, only one facial image was extracted from each video with the eye positions labeled manually. A more interesting problem is to look at how dynamics of the faces in the video can help improve the accuracy of FR. Although the algorithms from the participating institution were evaluated by the organizer, most of the datasets/partition information are not available online for reproducibility of the results.

3.2.3 "The Famous" Labeled Faces in the Wild, Its Old and New Protocols

There exist a large number of benchmark databases for evaluating FR algorithms, like FERET [15], AR [19] and ORL [83] just to name a few. A comprehensive list can be found in [84]. Most of the these databases are collected in controlled (studio) environment for studying certain aspects of FR (like expression and/or illumination) which are posed or unnatural. Under these controlled conditions, FR algorithms can achieve performance comparable to human beings. However, these algorithms cannot generalize well to data collected under different natural or spontaneous conditions. The LFW dataset [35] provides the FR community with uncontrolled face images from the web for pairwise matching/unmatching problem. The LFW dataset exhibits variability in lighting, pose, subject age, expression, race, gender and so on. The goals of the dataset are:

- Provide a large database of real world face images for the unseen pair matching problem of FR,
- Fit neatly into the detection-alignment-recognition pipeline, and
- Allow careful and easy comparison of FR algorithms.

The original LFW dataset contains 13,233 images of 5749 people, among which 1680 people have more than 1 image per person. The images were collected from online internet news articles and processed using Viola-Jones face detector [85] for detecting faces.

The Old Protocol

This dataset contains 300 pairs of genuine matches and 300 pairs of imposter matches for tenfold of cross validation leading to 3000 genuine and imposter matches each. The dataset is organized in 2 "views": view 1 is used for development training/testing purposes, where the training/test partitions are generated randomly and independently of the splits for tenfold of cross validation. This view is used for model selection and/or validation purposes. View 2 is used for performance testing and final evaluation of the algorithms to minimize fitting to the test data. View 2 is divided in ten subgroups such that the face pairs are mutually exclusive for tenfold of cross-validation, whose results are averaged to get the final performance of the model selected with view 1 data [86].

As running the Viola-Jones face detection algorithm [85] generated the face images, it fits well in the three-step detection-alignment-recognition pipeline for FR, (as explained in Sect. 2.2) and indeed, the latest LFW dataset includes four different sets of LFW images, the original and three different "aligned" images. The aligned versions include, (1) the "funneled images" (LFW-a) by Huang et al. [87], (2) for second version, an unpublished method is used for alignment of LFW-a [86] and (3) "deep funneled" images again by Huang et al. [88]. The last two funneled images produce superior results for most FV algorithms over the first two sets of images. From the evaluation of various algorithms, it is evident that the use of training data outside of LFW can have a significant impact on recognition performance. Numerous benchmarking results can be found in [46].

In conclusion, the LFW dataset provides the research community with a less controlled face dataset for FV system development. It has stimulated researchers to work on more "natural" and unconstrained FR problems that would generalize to data outside the existing dataset. However, as the face images were collected from news articles on the web, they are affected by the photographers' and editors' choice, so there were not many images under extreme lighting conditions. Since the faces are detected using the Viola-Jones detector, there are a limited number of faces with side views and views from above and below.

The New Challenging Protocol

If we think of a very common real-world scenario where 500,000 visitors visit an amusement park per day using facial biometrics, certain CVR at 0.1 % FAR implies that 500 people can falsely (with fake or shared identity) enter the park per day. This can be a big concern and loss to the park owner. Old LFW benchmark protocol contains 3000 pairs of genuine matches and 3000 pairs of imposter matches in total which are very limited to evaluate the large scale performance. Using old protocol, performance evaluation at false acceptance rate (FAR) of 0.1 % is not statistically significant as it requires to count only three imposters matching scores. A vast majority of researchers has been following this old protocol, that uses partial data

of this database to evaluate their algorithms (for details see LFW results website [46]). So there is a need to enhance the LFW benchmark protocol and exploit all the available data.

Liao et al. [6] designed a division of the LFW dataset into development-set that contains a set of training and testing data to tune the parameters. Also, an evaluation-set is designed to evaluate the performance of FR with 85,341 genuine matches, 6,122,185 imposter matches in training; 156,915 genuine matches and 46,960,863 imposter matches in testing. The new protocol takes into account large number of genuine and imposter matches both in the training and testing datasets and hence, it can evaluate very low FARs (e.g. <0.1 %), which are statistically significant.

Liao et al. [6] implemented seven learning techniques: PCA, LDA, large margin nearest neighbor (LMNN), information theoretical metric learning (ITML), keep it simple and straightforward metric learning (KISSME), locally-adaptive decision functions (LADF) and joint Bayesian formulation using three features namely, hand-crafted feature LBP, a learning based descriptor local embedding (LE) and high dimensional LBP (HighDimLBP) feature. FAR and open-set identification rates are measured as performance indicators. The best results are obtained using joint Bayesian approach with HighDimLBP features [89], where the CVR achieved is 41.66 % rates at FAR = 0.1 % and open-set identification rate as 18.07 % at rank 1 and FAR = 1 %. Therefore, it is evident from this recent benchmark study of large-scale unconstrained FR [6] that the newer protocol is very challenging and more practical as compared to the previously evaluated results [46].

Although this work added some improvements to the LFW benchmark study by increasing the number of correct and false matches obtained by the data, the CVR is too low to be considered in real-world FR applications. The performance is still far from satisfactory as the verification and identification rates are very poor under the large-scale unconstrained FR setting.

3.2.4 YouTube Video Database

LFW is a database used for evaluating FR algorithms with still face images recorded in uncontrolled conditions. As for videos, there exist several methods that have performed well in video FR tasks by exploiting the fact that a single face might appear in a video in consecutive frames [90, 91]. But the datasets used for developing those algorithms are primarily collected in highly controlled lighting and shooting conditions with high quality storage. In contrast, the YouTube face Dataset (YTF) [14] complements the LFW by providing a database of face videos designed for studying the problem of FR in videos with uncontrolled lighting, shooting condition and video quality. The videos were downloaded from YouTube with identities from the LFW dataset. Each video in YTF comes with a label indicating the identities of a person appearing in that video.

The dataset contains 3425 video clips of 1595 different people. The duration of these video clips ranges from 48 frames to 6070 frames with an average length of 181.3 frames per person. Because the videos were downloaded from

the YouTube using automatic tools, this dataset is highly uncontrolled in terms of lighting, shooting condition, video quality etc. Following the LFW protocol [35], the evaluation of algorithms on this dataset is a standard tenfold cross validation, pair-matching test. In the evaluation phase, 5000 video pairs are randomly selected form the dataset, in which half are matched pairs (same person) and half are unmatching pairs (different person). These pairs were divided into ten subgroups, each of which contains 250 matched pairs and 250 unmatched pairs. Each algorithm is trained on nine subsets and tested on the left 1 for 10 times with each of the subsets being the testing set once. The average performance is reported.

All video frames are encoded by several well-established image descriptors including LBP, center-symmetric LBP (CSLBP) and four-patch LBP. With these encodings, several types of methods have been evaluated with the YTF database. Because each video contains multiple frames and each frame can be encoded as a vector, the problem of matching the faces in a pair of videos becomes matching two sets of vectors. Three major groups of methods have been considered. The first group employs comparisons between pairs of face images from each of the two videos. The second group uses algebraic methods, which compare vector sets. A third group including the pyramid match kernel and the locality-constrained linear coding methods were effective in comparing sets of image descriptors. In total, the author of the dataset evaluated five groups of methods with three types of face image encoding and the results are shown in [14]. However, the best performance is reported using the DeepFace recording an accuracy of 91.4 %.

The contributions of the YouTube dataset and the evaluations include the following:

- A comprehensive dataset of labeled face videos in uncontrolled environment was presented together with benchmarks and pair-matching tests,
- The benchmark was used to compare a variety of existing video face matching methods and
- Stimulate further research in video FR in challenging and uncontrolled conditions.

3.3 Summary of the Emerging Databases

Four databases for FR have been discussed in the above subsections. These four databases together with six other emerging databases are summarized in Table 1. The FRVT 2013 provides with three main types of images for testing typical identity verification which could be deployed for detection of duplicates in databases, detection of fraudulent applications for credentials such as passports, criminal investigation, surveillance, and forensic clustering. The mugshot set and webcam set vary in their image quality and they can help study the effects of image quality on recognition performance. The evaluation also found that age of the people shown in the images also contribute to the performance of nearly all FR algorithms evaluated.

Table 1 Comparison of the emerging face recognition databases

Databases	Data source	Modality	Scenario	Data size	Subject size	Publicly available
LFW 2007 [6, 35]	Images from web news article	Still images	One-to-one FV & open-set FI	13,233	5749	Yes
GBU 2010 [69, 76]	Uncontrolled frontal images	Still images	One-to-one FV	1085 images in target & query sets × 3	437	No
YouTube 2011 [14]	Videos from YouTube	Videos	One-to-one FV	3425 clips	1595	Yes
MOBIO 2012 [80]	Collected with mobile phones	Videos with audio	One-to-many FI	61 h	152	Yes
Makeup database 2012 [92, 93]	Three categories from YouTube video makeup	Still female images	Makeup detection & one-to-one FV	604 + 204 + 154	151 + 51 + 124	Yes
FVRT 2013 [5]	Mugshot, webcam & visa images	Still images	Five tracks: one-to-one FV & FI, one-to-many FI, etc.	–	–	No
Indian movie face database 2013 [94]	Face images from movies	Still images	FI	34,512	100	Yes
McGillFaces database 2013 [95]	Indoor/outdoor uncontrolled face videos	Videos	Pose/gender/ face hair analysis	60 clips	60	Yes
Labeled wikipedia faces (LWF) 2014 [96]	Wikipedia biographic entries	Still images	One-to-one FV & FI	8500	1500	Yes
FaceScrub 2014 [97]	Public figures from searched queries	Still images	Detect target face from searched queries	107,818	530	Yes

Usually the older the people, the easier it is for the FR algorithm to recognize. Sketch faces based on FERET dataset was also used in the evaluation to support research in face sketch synthesis and recognition. The FRVT has provided a platform to test the commercial FR systems that have the potential to be deployed in different places by the US government and it also identities the future research directions for the FR research community.

GBU dataset provides three partitions of face images, each of different level of difficulty. The images were collected in a partially controlled environment where the pose and age are controlled but the expression, lighting are not. Because all faces are frontal faces, the only reason causing different recognition results is the representation of the faces in each image. FR algorithms can achieve better performance than humans in fully controlled condition [4]. While in fully uncontrolled conditions, no significant progress could be made. Thus the GBU dataset stimulates the development of robust frontal FR algorithms that could make progress in more challenging, partially controlled tasks without sacrificing its performance in easier ones.

The MOBIO face database consists of more than 61 h of audio-video bi-modal faces (also summarized in Table 1). The videos are recorded by handhold mobile phones, recording people speaking to the phone camera while answering a set of predefined questions. MOBIO provides the research community with a bi-modal dataset that could be used to evaluate speaker recognition, FR as well as their fusion. Since the videos are recorded by amateurs using mobile phones, there is large variability in pose, illumination, background environment as well as the audio-video quality. This nature of the dataset makes it challenging and encourages research to combine both modalities to improve the performance. However, in the evaluation stage, only individual frames containing faces were used to perform FR. A video based FR system should give better performance by exploiting the dynamics of the recorded faces. Another contribution of this dataset is to encourage the researchers to focus on the trade-off between performance and hardware requirement. Since the dataset is intended to stimulate development of algorithms that could find its applications in mobile devices, an important aspect of the evaluation to consider is the execution time and memory requirements.

Both LFW and YTF databases provide a large collection of faces recorded in uncontrolled conditions from the internet. For LFW, the face images are from online articles and each face comes with a label of the person's name. The YTF database takes a similar approach and the videos are downloaded from YouTube and also come with identity labels of the people. These two databases offer the FR community a good playground for developing and evaluating algorithms targeting at more natural and less controlled settings. For the LFW database, although the face images are more natural than those taken in fully controlled conditions, the images are often taken with good lighting and lack non-frontal faces.

In summary, from the benchmark databases presented above, we can see the following trend in FR research, benchmark database and protocol design:

1. As FR in controlled environment is considered a "solved" problem with some algorithms outperforming humans, the frontier of FR research is shifting to uncontrolled and more natural settings.
2. Coupled with powerful computing machines, improved algorithms for deep learning are able to discover patterns in large dataset. Hence larger labeled databases are desired in the FR community to develop large-scale and robust FR algorithms.
3. Nowadays, almost everyone cannot live without a mobile phone. FR systems on mobile phone and wearable devices would find its application in our everyday life. Thus robust FR algorithms running on mobile devices in natural settings will be of great value to the consumers.
4. As more and more algorithms are being developed for FR in videos, the dynamics of moving faces in videos should be further exploited to build more robust and accurate next generation FR systems.

From such papers, evaluations and benchmark competition results, it is apparent that unconstrained FR with large or small scale scenarios is largely an unsolved problem and should receive further attention. Human beings are amazing for FR under unconditioned settings. Even after years or with diverse makeups/appearences, human beings hardly fail to recognize an individual. Hence, it is imperative that we derive psychophysical and/or biological motivations from human beings on aspects that have made them experts in FR over centuries.

4 Human Recognition of Faces

The human face is perhaps the most important class of objects that we are interested to interact with. Our response to human faces is distinct from that to other classes of objects: there seemed to be a selective preference to human faces as we age. A study by Michael et al. [98] on 3-, 6-, 9-month old infants and adult groups revealed a greater percentage of gaze dwell time on faces with age. This selective attention of the human visual system towards other human faces might stem from having a default network in the brain that drives a series of involuntary cognitive processes: us thinking about recent events and speculating future ones that are founded on social interactions and involve the theory of mind [99] during periods of inactivity. Evidence from neuroimaging studies of brain diseases such as Alzheimer's, autism, schizophrenia depression etc., seemed to target and cripple this default network; therefore leading to the impairment of social cognitive abilities on varying degrees for patients with the aforementioned diseases.

The (hypothesized) existence of such network, one that attunes to social interactions and theory of mind, supports the fact that we gravitate towards connecting and understanding people above all others. This in turn, explains our selective preference to human faces and motivates the need to study how we perform the two main types of face recognition: face verification (for unfamiliar faces where the individual only

has a sense of familiarity of having seen the face before e.g. acquaintances) [100] and face identification (for familiar faces where the individual has both a sense of familiarity of having seen the face before and is able to identify him/her by name) as the foundation to successfully navigate the social world.

That being said, it will be beyond the scope of this chapter to involve all aspects of neuroscience, neuroimaging and psychological studies to explain the neuroanatomy of the default network and social cognition. Henceforth, the coverage of this part of the review will be dedicated to psychophysical and neuroimaging discoveries about the FR capabilities in humans. The sections that follow are the introduction to the two main hypotheses on the motion advantage in recognizing faces, with four other subsections on the current most difficult conditions pertaining to human performance in FR (the Big Four! [4])—illumination-, facial expression-, view perspective-, and age-invariant recognition. These four sections are distinct from the challenges (scale, occlusion and motion blur) in machine recognition of faces as discussed in Sect. 2. Finally, we would like to offer a preview of our integrated experimental approach that might be feasible in transcending FR across the four difficult conditions to achieve performance inspired by humans in an unconstrained and naturalistic setting.

4.1 Temporal Cues That Aid Face Recognition: Two Hypotheses to Explain Motion Advantage

Motion brings not only a face, but also the personality of its owner, to life. We are inherently dependent on the dynamics of motion to infer the mental states of those whom we are interacting in numerous social contexts. Visual inputs of the changes in head movements, varying degrees of facial expressions, eye gaze directions etc. bombard our senses in a myriad of signals before being integrated into a general, yet uncannily accurate, perception of how the present moment of interaction feels like. Such visual cues are essential in guiding our predictions of the 'appropriate' actions to take within a particular social context. If your counterpart is speaking to you while his eye gaze kept darting towards the nearest exit or his watch, you would probably have inferred that he is in a hurry to leave and that you should quickly wrap things up and end the conversation.

Similarly, motion in dynamic faces gives rise to a plethora of information that elevates both face verification and identification as compared to when static faces are presented. Hence, research on human face processing is now delving into dynamic faces to not only simulate a more realistic context for face recognition and processing, it is also an attempt at dissecting and comprehending how the presence of motion leads to improved face verification and identification performances. Consequently, two hypotheses were formulated in an effort to explain the benefits that dynamic faces impart on human recognition of faces: the supplemental information hypothesis and representation enhancement hypothesis [7].

The supplemental information hypothesis embodies an idea that the ventral temporal cortex, which includes the lateral fusiform gyrus, occipital face area and other associated structures in the human brain, is responsible for processing both invariant face information and the idiosyncratic facial motions of individuals. Inevitably, this confines the realization of the supplemental information hypothesis in FR to recognizing familiar faces only [7]. Work by Lander and Chuang [51] provided evidence that non-rigid facial motion (movement of internal facial features such as blinking of the eyes and chewing movements of the mouth), more than rigid motion (global movement of head including pan, tilt, yaw and other head translations), improves FI of familiar individuals. "Distinctive" facial motion (a separate entity from distinctive facial features) as well as "naturalness" of facial motion (not artificially designed motion) suggested in [101], proved facilitative to facial identification of familiar faces as well.

On the other hand, the representation enhancement hypothesis posits that recognition performance of a novel face after a learning phase has a higher accuracy than that learnt from static faces. This hypothesis is founded on a perspective-oriented learning of unfamiliar faces; also known as the "structure-from-motion" learning [102], where the advantage relies on the fact that knowledge about the three-dimensional structure of an individual's face can be gathered from motion prior to subsequent recognition. Such a learning process is said to confer humans the ability of recognizing unfamiliar faces. In an experiment by Pilz et al. [103], subjects were primed with unfamiliar faces with emotions of either a frown or a smile in non-frontal viewing perspectives and asked to do FV with a target face in the frontal view with an opposite emotion from that of the primed face. Results revealed that subjects generally responded faster when primed faces were in non-rigid motion as compared to static ones.

These are some of the psychophysical experiments that are seemingly representative of the two different hypotheses mentioned. They can be considered seminal works, which inspire later research to further refine FR experiments for the sake of allowing a better understanding of how this is done so effortlessly in humans. The following sections will discuss interesting findings for different facets of FR spanning from the various fields of study (psychophysics and neuroimaging) for cross validation and inspiration.

4.2 How Do Human Beings Handle the Big 4?

Similar to the challenges faced by computational models described in Sect. 2.1, we human beings also face difficulties in recognizing individuals across various conditions. In the following four subsections, we review some of the popular human centered experiments and protocols so as to understand innovative strategies, prior learning, biases at various levels and how exactly human beings overcome these four big problems.

4.2.1 Face Recognition Across Different Illumination Conditions

Astonishingly, there is little work done in terms of psychophysics experiments conducted to investigate how humans do FR across illumination variations. Instead, a plethora of work mostly centralized around improving or developing new pipelines for computer vision under this category.

Nonetheless, Tarr et al. [104] have shown that recognition performance for human is dependent on the difference between the degree of facial illumination presented to subjects during the training and testing phases. Subjects were first allowed to study a sheet of ten different frontal face images with their corresponding names (e.g. Allen, Laura) printed. The faces were shown illuminated from the front, normal to the face. For each face, the lighting space was sampled in 15° increments in both the horizontal and vertical axes to the right of the camera axis. In each of the five experiments conducted, subjects were shown different subsets of illuminated faces in the training phase (an illustration is shown in Fig. 14) before proceeding to the full set in the testing phase with large illumination variations (similar to the images shown in Fig. 1).

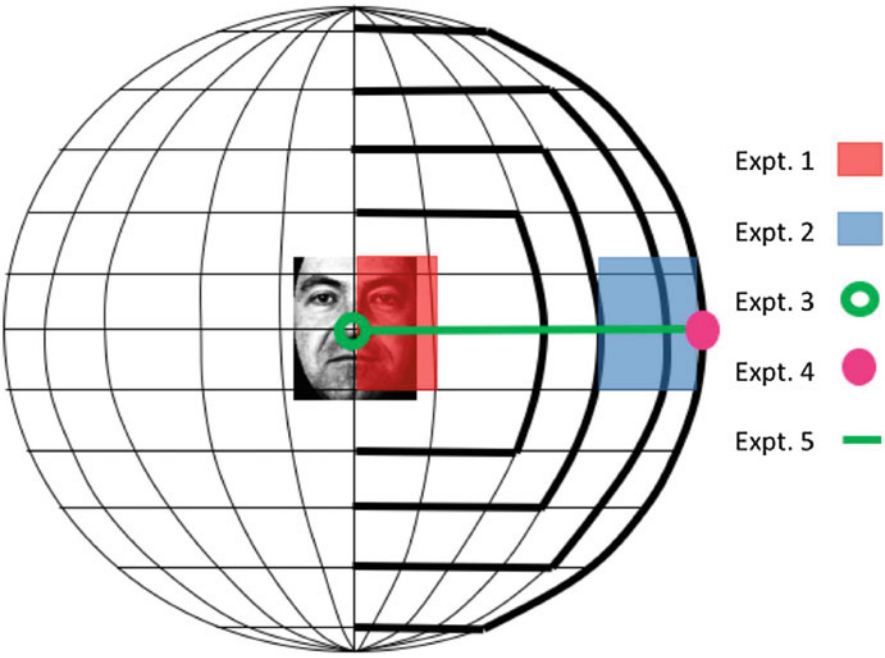

Fig. 14 Training sets for illumination variation experiments. Experiment 1 contains illuminations within 15° of the camera axis; Experiment 2 is a mirror of Experiment 1 with extreme lighting directions. Experiments 3 and 4 have one illumination condition each, (0°, 0°) and (75°, 0°) respectively. Experiment 5 contains illuminations along the horizontal meridian of the illumination space; from (0°, 0°) to (75°, 0°) (Best viewed in *color*)

Their results show that, in general, increasing the distance (i.e. the extent of difference) between illumination coordinates from the training and testing phases will decrease the FI performance of the subjects. Intuitively, we would expect performance to be worst for Experiments 2 and 4 (refer to Fig. 14), where subjects were trained with extreme illumination conditions ranging from 45° to 75° away from the normal. Yet, interestingly, the most prominent drop in performance was seen in Experiments 1, 3 and 5, where the face images were mostly illuminated from the frontal or near frontal coordinates. The authors reconciled this observation by explaining that because subjects were trained with extreme illumination conditions in Experiments 2 and 4, they were able to identify the faces with greater accuracies by using generic knowledge about geometry of faces as a class to infer their appearances under novel illumination circumstances. The ability of prediction can be attributed to the neural mechanisms of the posterior superior temporal sulcus (pSTS), which is responsible for processing changeable information in faces; where in this case it is the information on shape and surface orientations that is processed [105]. This could prove as evidence of the hypothesis that the dorsal stream pSTS identity representation might include a representation of facial shape that is independent of signature motions [102]. This experiment has shown that the humans are sensitive to the degree of face illumination conditions, and that learning from extreme degrees of illumination, albeit counter-intuitively, facilitates recognition of novel face configurations.

However, the experiment is still considered limited in terms of understanding how the human visual system actually compensate for dynamic variations in lighting. What the human visual system encodes is a continuum of illumination changes as the coordinates of the light source changes temporally, as opposed to the discreet increment of illumination changes in the experiment. There is, therefore, much to gather in terms of how the shape and geometry of an individual's face changes with illumination along the temporal dimension are encoded in the human visual system to confer us the high accuracy in FR under novel illumination conditions.

Sinha [106] revealed human psychophysical studies on a subset of the illumination spectrum of faces: contrast negation. It shows when concluding whether an image is a face, there is significant drop in performance with contrast negation.

As seen in Fig. 15, the patches in (a) and (b) have different overall brightness, but the images can still be discerned as illustrating the same object—a face. However, when comparing the patches in (a) and (c), where both have the same overall brightness, the object depicted may be perceived differently. It was concluded that the direction of brightness contrast, or otherwise known as contrast polarity, plays an important role in object perception and recognition.

Another study by Wallis et al. [107] using 3D face images confirmed the deduction that temporal cues in the context of varying illumination in motion functions like a 'perceptual glue' in human visual perception. Subjects showed the tendency to assume that they were viewing a single face sample when it was actually morphed to a different identity during the transition of varying illumination (refer to Fig. 16). The degree of this effect is influenced by the presence of a

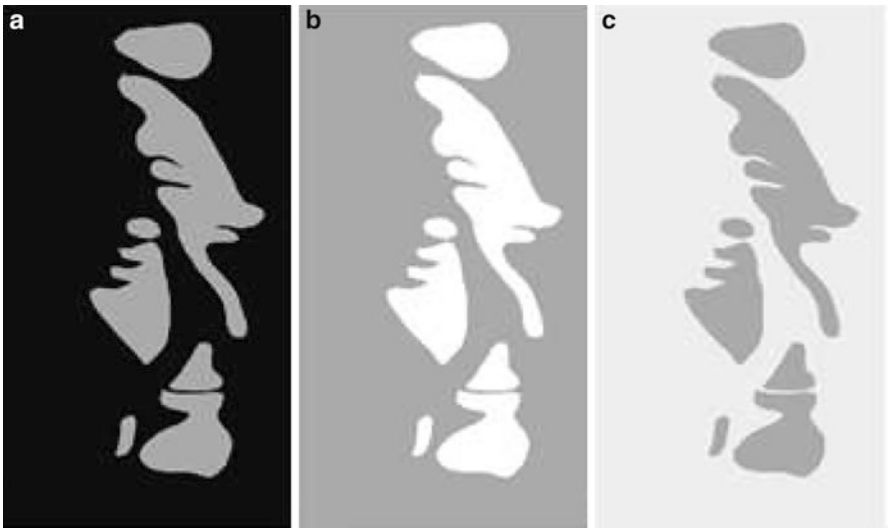

Fig. 15 Direction of contrast brightness affecting face detection, from Sinha [106]

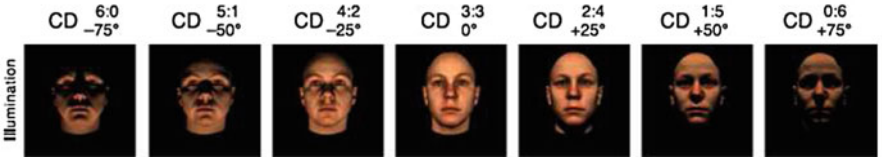

Fig. 16 Illumination varies during the morph transition from sample C to D in gradual ratio proportions, from Wallis et al. [107]

training phase where 'unlearning' the encoded visual representations of the sample is difficult, thereby supporting the representation enhancement hypothesis of the plausible "structure-from-motion" learning ability.

The context of any surface has an enormous effect on the color we see, e.g. illumination from the sun looks red in the evening, but yellow at noon and the recent 2015 debate on the color of a bodycon dress being blue-and-black or white-and-gold [108]. This is analogous to the scenario shown in Fig. 15. When there is a difference in contrast brightness and direction from the image's context to the object (i.e. both dark to light contrast seen in (a) and (b), but reversed for (c)), we see the conflation of illumination, reflectance and transmittance giving rise to the inverse optics problem; therefore leading to erroneous perception of the object [109, 110]. The peculiar way we see color and contrast, and hence, the way we perceive objects (especially faces) remains to be explained. Perhaps the answer to this problem is the way in which the human visual system copes with the inverse optics problem—a problem that could plausibly be simplified by investigating, on a frame-by-frame basis, how humans carry out FR across illumination variation leveraging on temporal dynamics in videos.

4.2.2 Face Recognition Across Different Facial Expressions

One of the most important skillsets for successful navigation in the social world is to accurately infer the mental and emotional states of others; hence our possession of a visual system that is attuned to human muscular motions [111]. One of the many such muscular motions is facial emotional expression—the main type of visual social cue that we infer from for information (such as the mental state of others, the intention behind their actions etc.) [112] on how appropriate we are to behave in a particular social context. Not only are facial expressions socially relevant, they are important in facilitating FI. The advantage for this aspect of FR is more profound when the facial expressions are presented in motion. Hereon, the main aspect of discussion will focus on the effects of emotional expressions on FR.

Firstly, there is the factor of idiosyncratic facial motion, which includes that of expressive (i.e. non-rigid motion) faces that promote higher FI accuracy. This is evident especially in the identification tasks involving faces that are familiar to the subjects during the experiment. The seminar work of [51] has shown that the accuracy for FI of familiar faces is the highest for expressing faces and that they outperform rigidly-moving (69.9 %, SD 14.5), talking (82.4 %, SD 11.7), and static faces (56.5 %, SD 22.0) at 89.5 % (SD 6.8) identification rate. Further investigation in a separate experiment showed 77.3 % (SD 12.2) of expressing faces possess distinctive facial movements during the course of expression and hence the high recognition rate for expressing faces in motion is obtained. This particular finding concurs with the supplemental information hypothesis and that there is a very strong motion advantage for identifying familiar faces based on a set of signature non-rigid facial movements for every individual. In other words, it can be argued that the processes for face expression and those for FI are integrated from plausibly different neural mechanisms in a manner that facilitates better performance in FR.

Other interesting results from two of their experiments is that there is a higher identification rate for familiar samples when the dynamics of non-rigid facial motion is natural; not artificially created or modified, and that the speed of the facial motion during expression is naturally fluid; not sped up or slowed intentionally [101]. An explanation for these behavioral phenomena can be found in recent neuroimaging data utilizing whole-brain analysis to show that the STS is the region with the greatest BOLD response under the influence of increased information in dynamic faces [113, 114] and fluidity of its motion [114]. What is especially interesting is that it reinforces the idea of distinct processing mechanisms devoted to facial identity and expression respectively. Majority of the ventral temporal face-sensitive regions of the brain (i.e. bilateral fusiform face area (FFA), occipital face area (OFA), right inferior occipital gyrus (IOG) and the right fusiform gyrus (FG)) seemed to be sensitive to the increased amount of frame information in dynamic faces, while a separate processing area is dedicated to the fluidity of that motion—STS [115]. Giese et al.'s [116] computational model of biological motion recognition has specified both a motion pathway and a form pathway in which neurons in the middle temporal (MT) area, the middle superior temporal (MST) area and the kinetic occipital (KO) area are attuned to discern optic flow localities before sending

its flow pattern to the STS for classification and identification in a feed-forward manner. This series of form detectors posits to supplement information from surface deformation of the face with the invariant face form learnt by the ventral temporal brain regions from multiple frames. Once again, this could potentially support the notion that enhanced FR from viewing the dynamics of natural facial expressions is not only dependent on the increased information presented in the form of increased number of frames, but relies on a disparate encoding process of an individual's non-rigid motion signature as well.

Another set of study by Rigby et al. [117] tested subjects on face processing where they have to make speeded expression (or identity) judgments of static and dynamic faces while identity (or expression) were held constant or varied. By showing that there was significant interference when processing static faces compared to dynamic faces, they provide evidence to support the idea that dynamic cues arising from the motion of facial expression, do facilitate a more efficient FI process. This dynamic advantage, however, was more obvious with the expression task as compared to the identity task. A plausible rationale behind such an observation could be that expressions causes global descriptors, which are crucial for holistic face processing, to be superseded by feature-based processing [28, 118]. Since facial expressions are considered socially salient and relevant [111], it is not surprising that humans will attend to and become adept at judging the types of expression.

Experiments on the composite effect of face processing, whereby feature-based face processing is dominant over holistic face processing, can serve as added evidence of having separate mechanisms for identity and expression recognition proficiency in humans. Underpinning this partial differentiation of the neuronal domains for identity and expression are human fMRI adaptation studies [119], as well as studies on prosopagnosic patients who possess the capability of recognizing facial expressions despite their disability in recognizing face identities [120, 121].

Concurring with this tenable anatomical discrimination of encoding for identity and expression, feature-based visual processing exuded in humans is demonstrated by Xiao et al. [122], who showed that subjects learning novel faces in non-rigid motion will have their feature-based FR less affected by irrelevant information in composite faces. They were more competent at verifying if the top and bottom halves of a face belonged to the same person after learning them in motion than with static faces. Such results are accounted for by the representation enhancement hypothesis—a 'structure-from-motion' type of learning. Perhaps, the motion information from the STS and other similarly committed brain regions is mapped in a piecewise manner to specific sites of the face according to the observed surface deformation in order for the brain to learn a set of signature facial movements for individuation. Using results from a classic experiment by Patterson and Baddeley [131] as an illustration, a simultaneous shift in both a face's viewing perspective and emotional expression between the learning and testing phases did not induce a significant drop in FI performance [123]. On the contrary, a sole change in viewing perspective will severely compromise FI performance during the testing phase. This suggests room for leveraging on the advantage of using dynamic facial expressions

to extrapolate recognition from neutral or a set of orthogonal expressions. After all, the act of smiling not only stretches one's mouth such that it takes up a larger area relative to a neutral face; yet it allows the viewer to visually encode the unique shape and trajectory of that smile to aid face discrimination. It, therefore, will be worthwhile to investigate the exact mechanism of mapping such a dynamic learning process to invariant recognition of faces.

4.2.3 Face Recognition Across Different Viewing Perspectives

Being immersed in a social world, we interact with people under unconstrained conditions on a daily basis. One of such conditions is the constant change in viewing perspective of a face. Be it listening to a presentation at a conference or talking to a group of friends, we all succumb to viewing faces in a range of different angles relative to the horizontal and vertical axes from the normal of the frontal view. Hence, it is relevant to understand how such rigid motion (e.g. yaw or pitch) contributes useful input to the human visual system for robust FR.

Intuitively, presentation of a face moving across different perspectives to a subject will provide him with more information of the overall 3D structure of the individual; but which aspects of a dynamic face allows for better FR? It was argued that perhaps the human visual system is evolved to achieve a representational structure that includes object information across both temporal and spatial dimensions.

In a FV task across different view perspectives, Pilz et al. [103] established evidence that learning a novel face in motion will lead to heightened FV performance, along with a shorter response time, as compared to that when learning a static face. This observation was obtained despite the fact that the target face to be matched was presented in a different viewing perspective from that of the learned dynamic face. Souza et al. [124] discovered a heterogeneous distribution of view-selective face neurons in the anterior STS (aSTS) that might be able to explain how humans learn face identity from different view angles. They found that in the caudal region of the aSTS, majority of the face-sensitive neurons elicited responses to the right and left views of a face. On the other hand, face-sensitive neurons in the rostral region showed a peak in response to a single oblique view only. This could imply that the processing of a face's different perspectives is conducted by having different populations of neurons represent specific sets of view angle information. Therefore, when testing a novel face's identity is conducted after learning from a dynamic face, a faster response time with higher recognition accuracy can plausibly be explained with the integration of view angle information gathered by neurons in the different regions of the aSTS, along with the identity information from the FG.

A lot of the image-based recognition of objects (including faces) is carried out at the level of fine abstract features [107]. Neurons learning the invariant properties of a feature not only capture information on the object's transformations, they might also generalize such learning to a diversity of objects that might contain the same feature. Such a theory may offer an explanation as to why humans are competent in identifying objects from novel viewing angles. It also functions as evidence that

temporal cues extracted from dynamic faces influence neural representations of objects by serving as the 'perceptual glue' to gel learnt concepts for subsequent recognition. However, the exact neural computation for such an abstract, generalized learning mechanism remains elusive to this day.

4.2.4 Face Recognition Across Age Differences

As humans age, the sands of time will etch a gradual, conspicuous trail of changes to the skin surface. Being a complex process, facial aging affects both the shape and texture of a face. Its manifestations vary among different age groups and ethnicities [125], with extrinsic factors like individual lifestyles and environmental conditions affecting the rate and extent of observable aging.

Understanding FR across time lapses in age is crucial especially in applications such as forensic art, electronic customer relationship management, security control and surveillance monitoring, biometrics, entertainment (e.g. accelerate actors' age in movies as required) and cosmetology [33]. Notwithstanding the relevance of FR in this aspect, there is very little work done using human psychophysics to study age invariant recognition in contrast to the vast amount of literature on computer vision techniques in this aspect.

Perhaps the computer can outperform human in terms of fine changes in facial such as the identification of craniofacial growth (i.e. changes in face shape) and the relative surface area and protrusion of facial features such as eyes, nose, ears, mouth, cheeks and chin (the cranium grows to cause sloping and shrinking of the forehead by releasing more space on the cranium's surface for those features) [126]. Skin texture will also be expected to change as collagen breaks down and the skin sags to form wrinkles due to its inability to maintain its former elasticity. At the same time, implications from previous exposure to the sun and age-related health problems like liver failure will begin to show in the form of hyper pigmentation patches and a yellowish complexion (e.g. jaundice) respectively on the skin's epidermal layers [127]. All these and more might not be as obvious to the human eye as it can be to a computer. However, given the premise that dynamic face information is omnipresent in reality, humans might be able to verify/identify a face with a much shorter processing time than machines. This age-invariant FR with motion is, unfortunately, not tested in any recent psychophysics or behavioral study for our evaluation. The closest we can get to human studies in this area of work is partially demonstrated by Suo et al. [128] where they did a simple human study after synthesizing new faces across different ages using a dynamic face aging model with multi-resolution and multi-layer image representations.

Given that their experiment is computational in nature (refer to Fig. 17 as an illustration) for an outline of their pipeline, the purpose of the included human study was to validate that their dynamic face aging model approximates to human perception. They did so in two separate experiments: one required the subjects to give estimates for the age of the face seen in original images and those generated by their model. The other task required subjects to match synthesized images of aged

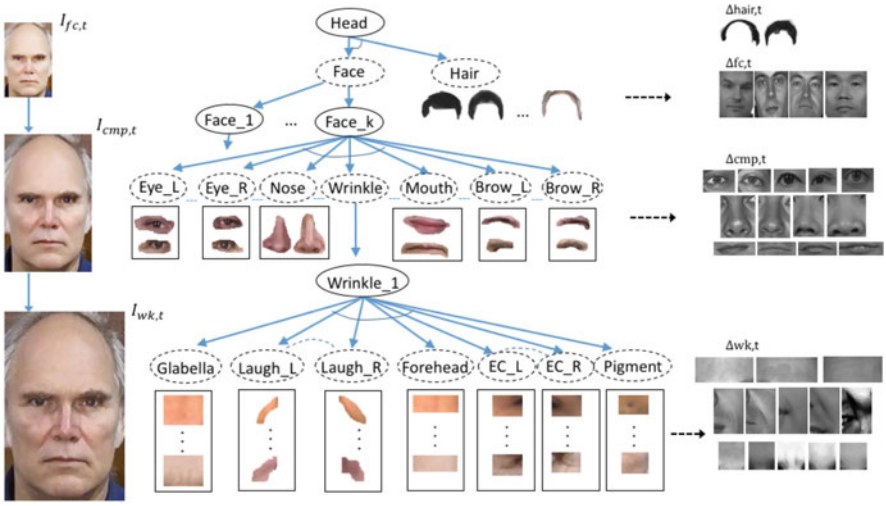

Fig. 17 Brief pipeline for dynamic face aging model

faces to their corresponding 'younger' face images. While subjects' estimates in the first task were consistently precise with the age of synthesized images, they did not perform as well in the second face-matching task. It was shown that identification performance decreases as the age difference between synthesized and real images increases.

Predicting how a person will look like in future as he/she ages is a hard task since everyone does so at different rates and that the signs of aging will differ from person to person. Moreover, using static images will be reducing FV to a picture-matching task that does not reflect the practical circumstances in which humans do so naturally—seeing faces in motion. It is plausible that certain facial features, given learning with motion, could serve as cues for invariant recognition across age (e.g. unique face surface deformations when smiling or frowning).

4.3 Transcending the Big Four: Evaluating Human Performance in Dynamic Perspective Invariant Face Recognition

We understand that prior work suggested that dynamic motion provides additional information, given an increase in number of frames to the identity of a face than static images, for an efficient FR system that allows human to navigate successfully in a social world. What remains unknown thus far is the type of additional information that can be gleaned from motion to support this inherent capability of mankind. In addition, the design of one or more FR tasks do not

mimic the complex ones which we face in the natural world: recognizing a friend in a mall or an unfamiliar keynote speaker at a conference meeting when he/she is conversing with another (i.e. combination of changing view perspective with expressive motion), recognizing a relative whom you've not seen in years as he/she darts into a sheltered building on a sunny mid-afternoon (i.e. combination of view perspective illumination and age variations) etc. Hence, there exist the knowledge gap as to how humans tackle such challenging recognition circumstances (situations where several conditions are confounded) seamlessly and effortlessly.

With well-established work done with dynamic FR, we question if there exist a generic strategy for each type and/or combination of conditions employed by humans given unconstrained viewing of faces in motion. Some might reckon that machine FR performance could very well have surpassed that of human's [2], with a 97 % correction recognition rate by the standard LFW protocol. However, it entails a 3 % FAR (False Acceptance Rate) that is unsatisfactory for practical applications. Even at 0.1 % FAR, the algorithm cannot be implemented for large-scale recognition as discussed in Sect. 3.2.3, e.g. airport security which handles hundreds of thousands of people daily (large number of genuine and imposter matches) [6].

Therefore, we design our experiment to investigate the plausible facial features and eye-gaze strategies of previously unfamiliar faces to be learnt in dynamic motion for subsequent recognition tasks [129]. This psychophysical experiment to be conducted aims to obtain inspirations from highly-competent human subjects to determine if generic eye gaze scan path strategies, as well as crucial facial features, can be used to explain FV for the different realistic scenarios occurring around us on an everyday basis.

Subjects will be presented with pairs of dynamic face samples recorded in an array of different unconstrained settings and they will be asked to identify if the two faces belonged to the same person (i.e. FV). Key features from the tests can then be evaluated so that we may emulate the competence of the human recognition system to push the boundaries of machine FR.

5 Summary and Future Trends

Rigorous and huge amount of research efforts from diverse fields of studies like computer, cognitive and biological sciences, are aiming to tame the challenging problems of FR. As we have seen over the period of years, for constrained and well-conditioned limited cases, the field of FR has reached a certain level of maturity. However, a vast majority of unconstrained FR cases require further attention and new directions in their investigations. FR using videos is going to play a much bigger role in the years to come. As explained in Sect. 2.3.2, with hardware devices for computing, recording and storing the relevant data are becoming cheaper and more readily available, people will be able to perform their vision based tasks (such as FR) in video-to-video scenarios. The rich temporal information available in such modalities (captured under scenarios described in Sect. 2.1) makes it very appealing

and attractive to many researchers. So it should be really exciting and challenging for researchers to find new methodologies for video based FR involving very high, large scale face voluminous data.

In recent times, DNN involved in deep learning architecture based methodologies using gigantic amount of training face images and hundreds of millions of parameters have shown surprisingly outstanding results. Results shown by DeepFace, DeepID and few others, are really impressive and they outperform most of the handcrafted features obtained using traditional machine learning approaches. However, their evaluations on large scale unconstrained FR problem as described in Sect. 3.2.3 are yet to be done. Moreover, their training images, architecture and learning frameworks are proprietary, thereby leaving very limited scopes for further research using large scale training images. One important factor that researchers can look into is how to develop DNN framework using lower number of training samples and how biologically inspired networks can be incorporated in DNN framework.

On the human FR aspect, researchers in cognitive science are moving away from how humans recognize still face images to recognition of faces in videos [130]. As explained in Sect. 4.1, the formulation of two hypotheses by O'Toole et al., is an attempt to explain the benefits that dynamic faces impart on human recognition of faces. The supplemental information hypothesis asserts FR depends on the representation of features and/or motion that is unique to an individual—his facial identity signature. Most experiments using familiar faces as stimuli fall under this category. Lander and Chuang showed that facial motion, specifically non-rigid motion, improves identification of faces when the dynamics are labeled 'distinctive', possess 'naturalness' (i.e. no artificial animations) in the motion and are viewed at naturalistic speeds.

On the other hand, the representation enhancement hypothesis posits that recognition performance of a novel face after a learning phase has a higher accuracy than that learnt from static faces. Lander and Bruce experiments have shown a heightened recognition performance after learning an unfamiliar face in motion. Although some argued that multiple static images of a face in different perspectives might be able to account for such learning advantage, studies by Pike et al. suggests that performance worsened when faces are learnt using a series of static images viewed in random order. These, and a few other works, showed that the dynamics of faces provide the viewer with a 3D structure that cannot be derived from multiple static views alone; hence making the study of faces in motion attractive.

The emerging directions discussed in Sect. 2.3 shed some light on how researchers are making fresh efforts in alleviating FR problems. One of the areas that has received attention is the biologically motivated approaches for FR. Understanding the invariance identity-preserving transformation theory may help to extract features that are invariant to certain transformations (may not be all). This would help to completely eradicate pre-processing stages in FR pipelines when processing raw images; thereby increasing their computational efficiencies and reducing the error rates at various stages.

Understanding how humans perform FR via behavioral studies can provide the first peek as to how the human brain gleans information from the external world in

which we interact. This is the motivation driving our experiment: the first step to emulate naturalistic FR. However, the research sphere for FV is relatively nucleated as compared to FI; with the latter involving information association and retrieval from the human memory. As such, we propound the notion of a two-prong approach to investigate the human memory and human performance in FR to plausibly leapfrog the long standing hurdles of machine recognition of faces.

References

1. M. Turk, A. Pentland, Eigenfaces for recognition. J. Cogn. Neurosci. **3**(1), 71–86 (1991)
2. Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1701–1708
3. Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1891–1898
4. P.J. Phillips, Face & Ocular Challenges. Presentation (2010), http://www.cse.nd.edu/BTAS_10/BTAS_Jonathon_Phillips_Sep_2010_FINAL.pdf
5. P. Grother, M. Ngan, Face Recognition Vendor Test (FRVT 2013) performance of face identification algorithms. Technical Report (2013), http://www.biometrics.nist.gov/cs_links/face/frvt/frvt2013/NIST_8009.pdf
6. S. Liao, Z. Lei, D. Yi, S.Z. Li, A benchmark study of large-scale unconstrained face recognition, in *IEEE International Joint Conference on Biometrics*, Clearwater, FL, 2014, pp. 1–8
7. A.J. O'Toole, D.A. Roark, H. Abdi, Recognizing moving faces: a psychological and neural synthesis. Trends Cogn. Sci. **6**(6), 261–266 (2002)
8. W.A. Bainbridge, P. Isola, A. Oliva, The intrinsic memorability of face photographs. J. Exp. Psychol. Gen. **4**(142), 1323–1334 (2013)
9. T.A. Busey, Formal models of familiarity and memorability in face recognition, in *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*, ed. by M.J. Wenger, J.T. Townsend (Lawrence Erlbaum Associates Publishers, Mahwah, 2001)
10. S. Georghiades, P.N. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 643–660 (2001)
11. L. Zhang, D. Samaras, Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. IEEE Trans. Pattern Anal. Mach. Intell. **28**(3), 351–363 (2006)
12. S. Vural, Y. Mae, H. Uvet, T. Arai, Illumination normalization for outdoor face recognition by using ayofa-filters. J. Pattern Recognit. Res. **6**(1), 1–18 (2011)
13. X. Zhao, S.K. Shah, I.A. Kakadiaris, Illumination alignment using lighting ratio: application to 3D-2D face recognition, in *Proceedings of International Conference on Automatic Face Gesture Recognition*, Shanghai, 2013, pp. 1–6
14. L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained video with matched background similarity, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, 2011, pp. 529–534
15. P.J. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
16. The Face Recognition Technology (FERET) Normalization (2005), http://www.cs.colostate.edu/evalfacerec/data/normalization.html

17. C. Ding, D. Tao, A comprehensive survey on pose-invariant face recognition. CoRR abs/1502.04383 (2015), http://www.arxiv.org/abs/1502.04383
18. X. Zhang, Y. Gao, Face recognition across pose: a review. Pattern Recogn. **42**, 2876–2896 (2009)
19. A.M. Martinez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. Pattern Anal. Mach. Intell. **24**(6), 748–763 (2002)
20. B. Mandal, X.D. Jiang, A. Kot, Verification of human faces using predicted eigenvalues, in *19th International Conference on Pattern Recognition*, Tempa, FL, 2008, pp. 1–4
21. J. Leibo, Q. Liao, T. Poggio, Subtasks of unconstrained face recognition, in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, Lisbon, vol. 2, 2014, pp. 113–121
22. P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005, pp. 947–954
23. B. Mandal, W. Zhikai, L. Li, A. Kassim, Evaluation of descriptors and distance measures on benchmarks and first-person-view videos for face identification, in *International Workshop on Robust Local Descriptors for Computer Vision*, Singapore, 2014, pp. 585–599
24. X.D. Jiang, B. Mandal, A. Kot, Eigenfeature regularization and extraction in face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**(3), 383–394 (2008)
25. M. Kawulok, J. Szymanek, Precise multi-level face detector for advanced analysis of facial images. IET Image Process. **6**(2), 95–103 (2012)
26. C. Zhang, Z. Zhang, A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, http://www.research.microsoft.com/pubs/132077/facedetsurvey.pdf
27. B. Mandal, S. Ching, L. Li, V. Chandrasekha, C. Tan, J.-H. Lim, A wearable face recognition system on Google glass for assisting social interactions, in *Third International Workshop on Intelligent Mobile and Egocentric Vision*, Singapore, 2014, pp. 419–433
28. W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey. ACM Comput. Surv. **35**(4), 399–458 (2003)
29. B. Klare, A. Jain, On a taxonomy of facial features, in *Proceedings of International Conference on Biometrics: Theory, Applications and Systems* (2010), pp. 1–8
30. H. Zhou, A. Mian, L. Wei, D. Creighton, M. Hossny, S. Nahavandi, Recent advances on singlemodal and multimodal face recognition: a survey. IEEE Trans. Hum. Mach. Syst. **44**(6), 701–716 (2014)
31. P. Belhumeur, Ongoing challenges in face recognition, in *Frontiers of Engineering: Reports on Leading-Edge Engineering* (2006), pp. 5–14
32. X. Zou, J. Kittler, K. Messer, Illumination invariant face recognition: a survey, in *Proceedings of International Conference on Biometrics: Theory, Applications, and Systems* (2007), pp. 1–8
33. Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **32**(11), 1955–1976 (2010)
34. A. Jain, B. Klare, U. Park, Face recognition: some challenges in forensics, in *Proceedings of International Conference on Automatic Face Gesture Recognition and Workshops* (2011), pp. 726–733
35. G. Huang, M. Ramesh, T. Berg, E.L. Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07-49 (University of Massachusetts, Amherst, 2007)
36. T. Poggio, J. Mutch, F. Anselmi, J. Leibo, L. Rosasco, A. Tacchetti, The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work). MIT-CSAIL-TR-2012-035
37. Q. Liao, J. Leibo, T. Poggio, Learning invariant representations and applications to face verification, in *Neural Information Processing Systems Foundation, Inc.*, Harrahs and Harveys, Lake Tahoe, USA (2013), pp. 1–9
38. S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*, 2nd edn. (Springer, Berlin, 2011)

39. J. Barr, K. Bowyer, P. Flynn, S. Biswas, Face recognition from video: a review. Int. J. Pattern Recognit. Artif. Intell. **26**(5), (2012), DOI: 10.1142/S0218001412660024
40. F. Anselmi, J. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio, Unsupervised learning of invariant representations in hierarchical, architectures. arXiv preprint arXiv:1311.4158 (2013)
41. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. Nat. Neurosci. **2**(11), 1019–1025 (1999)
42. T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 411–426 (2007)
43. Y. LeCun, Y. Bengio (eds.), Convolutional networks for images, speech, and time series, in *The Handbook of Brain Theory and Neural Networks*, ACM Digital Library (1995)
44. Q. Liao, J.Z. Leibo, Y. Mroueh, T. Poggio, Can a biologically-plausible hierarchy effectively replace face detection, alignment and recognition pipelines? arXiv:1311.4082v3 [cs.CV], no. 003 (2013)
45. E. Hubel, T. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. **160**(1), 106 (1962)
46. Labeled faces in the wild (LFW) results (2015), http://www.vis-www.cs.umass.edu/lfw/results.html
47. L. Rowden, B. Klare, J. Klontz, A.K. Jain, Video-to-video face matching: establishing a baseline for unconstrained face recognition, in *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems*, Washington, DC, 2013, pp. 1–8
48. Instagram, Online social network through images (2015), https://www.instagram.com/
49. Facebook, Online social network (2015), https://www.facebook.com/
50. A. Ishai, L.G. Ungerleider, A. Martin, J.L. Schouten, J.V. Haxby, Distributed representations of objects in the human ventral visual pathway. Proc. Natl. Acad. Sci. **96**, 9379–9384 (1999)
51. K. Lander, L. Chuang, Why are moving faces easier to recognize? Vis. Cogn. **12**(3), 429–442 (2005)
52. W.A. Bainbridge, P. Isola, I. Blank, A. Oliva, Establishing a database for studying human face photograph memory, in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Austin, TX, 2012, pp. 1302–1307
53. Face Recognition Homepage Databases (2015), http://www.face-rec.org/databases/
54. U. Park, A.K. Jain, A. Ross, Face recognition in video: adaptive fusion of multiple matchers, in *IEEE Computer Workshop on Biometrics*, Minneapolis, 2007, pp. 1–8
55. H. Li, G. Hua, X. Shen, Z. Lin, J. Brandt, Eigen-PEP for video face recognition, in *Asian Conference on Computer Vision*, Singapore, 2014, pp. 17–33
56. Y. Chen, V. Patel, S. Shekhar, R. Chellappa, P. Phillips, Video-based face recognition via joint sparse representation, in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Shanghai, 2013, pp. 1–8
57. R. Chellappa, J. Ni, V. M. Patel, Remote identification of faces: problems, prospects, and progress. Pattern Recogn. Lett. **33**(15), 1849–1859 (2012)
58. B. Mandal, X.D. Jiang, A. Kot, Dimensionality reduction in subspace face recognition, in *Sixth IEEE International Conference on Information, Communications & Signal Processing*, Singapore, 2007, pp. 1–5
59. X.D. Jiang, B. Mandal, A. Kot, Complete discriminant evaluation and feature extraction in kernel space for face recognition. Mach. Vis. Appl. (Springer) **20**(1), 35–46 (2009)
60. B. Mandal, H. Eng, Regularized discriminant analysis for holistic human activity recognition. IEEE Intell. Syst. **27**(1), 21–31 (2012)
61. X.D. Jiang, B. Mandal, A. Kot, Enhanced maximum likelihood face recognition. IEE Electron. Lett. **42**(19), 1089–1090 (2006)
62. Z. Zhu, P. Luo, X. Wang, X. Tang, Deep learning identity preserving face space, in *International Conference on Computer Vision*, Washington, DC, 2013, pp. 113–120
63. Y. Sun, X. Wang, X. Tang, Hybrid deep learning for face verification, in *IEEE International Conference on Computer Vision*, Sydney, 2013, pp. 1489–1496
64. 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)

65. Face Recognition Technology (FERET) (1996), http://www.nist.gov/itl/iad/ig/feret.cfm
66. Face Recognition Vendor Test (FRVT) (2015), http://www.nist.gov/itl/iad/ig/frvt-home.cfm
67. Face Recognition Grand Challenge (FRGC) (2005), http://www.nist.gov/itl/iad/ig/frgc.cfm
68. Multiple Biometric Grand Challenge (MBGC) (2009), http://www.nist.gov/itl/iad/ig/mbgc.cfm
69. Face and Ocular Challenge Series (FOCS): Good, Bad and the Ugly Database (2015), http://www.nist.gov/itl/iad/ig/focs.cfm
70. P. Phillips, J. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y.M. Lui, H. Sahibzada, S. Weimer, An introduction to the good, the bad and the ugly face recognition challenge problem, in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops* (2011), pp. 346–353
71. Multiple Biometrics Evaluation (MBE) (2010), http://www.nist.gov/itl/iad/ig/mbe.cfm
72. Competition on Face Recognition in Mobile Environment (MOBIO) (2013), https://www.idiap.ch/dataset/mobio
73. Point and Shoot Face Recognition Challenge (PaSC) (2015), http://www.nist.gov/itl/iad/ig/pasc.cfm
74. X. Wang, X. Tang, Face photo-sketch synthesis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(11), 1955–1967 (2009)
75. W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, 2011, pp. 513–520
76. P. Grother, G.W. Quinn, P.J. Phillips, MBE 2010: report on the evaluation of 2D still-image face recognition algorithms. National Institute of Standards and Technology, NISTIR 7709
77. P.J. Phillips, W.T. Scruggs, A.J. OToole, P.J. Flynn, K.W. Bowyer, C.L. Schott, M. Sharpe, FRVT 2006 and ICE 2006 large-scale results. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 831–846 (2010)
78. Face Recognition Vendor Test (FRVT2006) (2006), http://www.nist.gov/itl/iad/ig/frvt-2006.cfm
79. A. O'Toole, P. Phillips, X. An, J. Dunlop, Demographic effects on estimates of automatic face recognition performance, in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, Santa Barbara, CA 2011, pp. 83–90
80. C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, T. Cootes, Bi-modal person recognition on a mobile phone: using mobile phone data, in *IEEE International Conference on Multimedia and Expo Workshops*, 2012, pp. 635–640
81. M. Gunther et al., Face recognition evaluation in mobile environment, in *International Conference on Biometrics*, Madrid, 2013, pp. 1–7
82. N. Pinto, D. Cox, Beyond simple features: a large-scale feature search approach to unconstrained face recognition, in *IEEE Automatic Face and Gesture Recognition*, Santa Barbara, CA, 2011, pp. 8–15
83. F. Samaria, A. Harter, Parameterization of a stochastic model for human face identification, in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, 1994, pp. 138–142
84. Face Recognition Databases (2015), http://www.face-rec.org/databases/
85. P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1 (2001), pp. 511–518
86. Labeled Faces in the Wild (LFW) (2015), http://www.vis-www.cs.umass.edu/lfw/
87. G. Huang, V. Jain, Unsupervised joint alignment of complex images, in *IEEE International Conference on Computer Vision*, Rio de Janeiro, 2007, pp. 1–8
88. G. Huang, M. Mattar, H. Lee, E.G. Learned-Miller, Learning to align from scratch, in *Advances in Neural Information Processing Systems*, vol. 25 (2012), pp. 764–772
89. D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: a joint formulation, in *European Conference on Computer Vision*, Florence, 2012, pp. 566–579

90. M. Everingham, J. Sivic, A. Zisserman, Taking the bite out of automated naming of characters in tv video. Image Vis. Comput. **27**(5), 545–559 (2009)
91. D. Ramanan, S. Baker, S. Kakade, Leveraging archival video for building face datasets, in *IEEE International Conference on Computer Vision* (2007), pp. 1–8
92. A. Dantcheva, C. Chen, A. Ross, Can facial cosmetics affect the matching accuracy of face recognition systems? in *2012 IEEE 5th International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (2012), pp. 391–398
93. C. Chen, A. Dantcheva, A. Ross, Automatic facial makeup detection with application in face recognition, in *International Conference on Biometrics* (IEEE, Madrid, 2013), pp. 1–8
94. S. Setty et al., Indian movie face database: a benchmark for face recognition under wide variations, in *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, Jodhpur, 2013, pp. 726–733
95. M. Demirkus, J.J. Clark, T. Arbel, Robust semi-automatic head pose labeling for real-world face video sequences. Multimedia Tools Appl., **70**(1), 495–523 (2014)
96. M.K. Hasan, C.J. Pal, Experiments on visual information extraction with the faces of wikipedia, in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 27–31 July 2014, Québec City, QC, 2014, pp. 51–58
97. H.-W. Ng, S. Winkler, A data-driven approach to cleaning large face datasets, in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 343–347
98. M.C. Frank, E. Vul, S.P. Johnson, Development of infants' attention to faces during the first year. Cognition **110**(2), 160–170 (2009)
99. R.L. Buckner, The serendipitous discovery of the brain's default network. NeuroImage **62**(2), 1137–1145 (2012)
100. B. Mandal, X. Jiang, H. Eng, A. Kot, Prediction of eigenvalues and regularization of eigenfeatures for human face verification. Pattern Recogn. Lett. **31**(8), 717–724 (2010)
101. K. Lander, L. Chuang, L. Wickham, Recognizing face identity from natural and morphed smiles. Q. J. Exp. Psychol. **59**(5), 801–808 (2006)
102. A. O'Toole, D. Roark, *Dynamic Faces: Memory for Moving Faces* (The MIT Press, Cambridge, 2011)
103. K.S. Pilz, I.M. Thornton, H.H. Bülthoff, A search advantage for faces learned in motion. Exp. Brain Res. **171**(4), 436–447 (2005)
104. M.J. Tarr, A.S. Georghiades, C.D. Jackson, Identifying faces across variations in lighting: psychophysics and computation. ACM Trans. Appl. Percept. **5**(2), 10:1–10:25 (2008)
105. J.V. Haxby, E.A. Hoffman, M.I. Gobbini, The distributed human neural system for face perception. Trends Cogn. Sci. **4**(6), 223–233 (2000)
106. P. Sinha, *Qualitative Representations for Recognition* (Springer, Berlin, 2002), pp. 249–262
107. G. Wallis, B.T. Backus, M. Langer, Learning illumination- and orientation-invariant representations of objects through temporal associations. J. Vis. **9**(7), 1–8 (2009)
108. The Science of Why No One Agrees on the Color of This Dress (2015), http://www.wired.com/2015/02/science-one-agrees-color-dress
109. D. Purves, *Brains: How They Seem to Work* (Pearson, Financial Times Press, New York, 2010)
110. D. Purves, R. Cabeza, S.A. Huettel, K.S. LaBar, M.L. Platt, M.G. Woldorff, *Principles of Cognitive Neuroscience*, 2nd edn., Sunderland, MA 01375-0407, USA, (Sinauer Associates, 2012)
111. R. Blake, M. Shiffrar, Perception of human motion. Annu. Rev. Psychol. **58**, 47–73 (2007)
112. M. Kamachi, V. Bruce, S. Mukaida, J. Gyoba, S. Yoshikawa, S. Akamatsu, Dynamic properties influence the perception of facial expressions. Perception **30**(7), 875–887 (2001)
113. D. Pitcher, D.D. Dilks, R.R. Saxe, C. Triantafyllou, N. Kanwisher, Differential selectivity for dynamic versus static information in face-selective cortical regions. NeuroImage **56**(4), 2356–2363 (2011)
114. J. Schultz, M. Brockhaus, H.H. Bülthoff, K.S. Pilz, What the human brain likes about facial motion. Cereb. Cortex **23**, 1167–1178 (2012)
115. C.P. Said, J.V. Haxby, A. Todorov, Brain systems for assessing the affective value of faces. Philos. Trans. R. Soc. B **366**, 1660–1670 (2011)

116. M.A. Giese, T. Poggio, Neural mechanisms for the recognition of biological movements. Nat. Rev. Neurosci. **4**, 179–192 (2003)
117. S. Rigby, B. Stoesz, L. Jakobson, How dynamic facial cues, stimulus orientation and processing biases influence identity and expression interference. J. Vis. **13**(9), 413–418 (2013)
118. X.D. Jiang, B. Mandal, A. Kot, Face recognition based on discriminant evaluation in the whole space, in *IEEE 32nd International Conference on Acoustics, Speech and Signal Processing*, Honolulu, Hawaii, 2007, pp. 245–248
119. J.S. Winston, R. Henson, M.R. Fine-Goulden, R.J. Dolan, fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. J. Neurophysiol. **92**(3), 1830–1839 (2004)
120. S. Bentin, J.M. DeGutis, M. D'Esposito, L.C. Robertson, Too many trees to see the forest: Performance, event-related potential, and functional magnetic resonance imaging manifestations of integrative congenital prosopagnosia. J. Cogn. Neurosci. **19**(1), 132–146 (2007)
121. B.C. Duchaine, H. Parker, K. Nakayama, Normal recognition of emotion in a prosopagnosic. Perception **32**, 827–838 (2003)
122. N.G. Xiao, P.C. Quinn, L. Ge, K. Lee, Elastic facial movement influences part-based but not holistic processing. J. Exp. Psychol. Hum. Percept. Perform. **39**(5), 1457–1467 (2013)
123. M.T. Posamentier, H. Abdi, Processing faces and facial expressions. Neuropsychol. Rev. **13**(3), 113–143 (2003)
124. W.C.D. Souza, S. Eifuku, R. Tamura, H. Nishijo, T. Ono, Differential characteristics of face neuron responses within the anterior superior temporal sulcus of macaques. J. Neurophysiol. **94**, 1252–1266 (2005)
125. U. Park, Y. Tong, A.K. Jain, Age-invariant face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 947–954 (2010)
126. B. Mandal, X.D. Jiang, A. Kot, Multi-scale feature extraction for face recognition, in *IEEE International Conference on Industrial Electronics and Applications*, Singapore, 2006, pp. 1–6
127. T. Igarashi, K. Nishino, S.K. Nayar, The appearance of human skin. Technical Report, Columbia University, New York (2005)
128. J. Suo, F. Min, S. Zhu, S. Shan, X. Chen, A multi-resolution dynamic model for face aging simulation, in *IEEE Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007, pp. 1–8
129. R. Lim, M.R. Sayed, B. Mandal, K.T. Ma, L. Li, J.H. Lim, Evaluating human performance in dynamic perspective invariant face recognition (accepted), in *11th Asia-Pacific Conference on Vision*, Singapore, 2015
130. G. Kreiman, Kreiman's lab (2015), http://www.klab.tch.harvard.edu/publications/publications.html
131. K. Patterson, A.D. Baddeley, When face recognition fails. J. Exp. Psychol. Hum. Learn. Mem. **3**(4), 406–417 (1977)