Visual Psychophysics for Making Face Recognition Algorithms More Explainable

Brandon RichardWebster[1](✉), So Yon Kwon[2], Christopher Clarizio[1], Samuel E. Anthony[2,3], and Walter J. Scheirer[1]

[1] University of Notre Dame, Notre Dame, IN 46556, USA
brichar1@nd.edu
[2] Perceptive Automata, Inc., Somerville, USA
[3] Harvard University, Cambridge, MA 02138, USA

Abstract. Scientific fields that are interested in faces have developed their own sets of concepts and procedures for understanding how a target model system (be it a person or algorithm) perceives a face under varying conditions. In computer vision, this has largely been in the form of dataset evaluation for recognition tasks where summary statistics are used to measure progress. While aggregate performance has continued to improve, understanding individual causes of failure has been difficult, as it is not always clear why a particular face fails to be recognized, or why an impostor is recognized by an algorithm. Importantly, other fields studying vision have addressed this via the use of visual psychophysics: the controlled manipulation of stimuli and careful study of the responses they evoke in a model system. In this paper, we suggest that visual psychophysics is a viable methodology for making face recognition algorithms more explainable. A comprehensive set of procedures is developed for assessing face recognition algorithm behavior, which is then deployed over state-of-the-art convolutional neural networks and more basic, yet still widely used, shallow and handcrafted feature-based approaches.

Keywords: Face recognition · Biometrics · Explainable AI
Visual psychophysics · Biometric menagerie

1 Introduction

With much fanfare, Apple unveiled its Face ID product for the iPhone X in the Fall of 2017 at what was supposed to be a highly scripted event for the media. Touted as one of the most sophisticated facial recognition capabilities available to consumers, Face ID was designed to tolerate the wide range of user behaviors and environmental conditions that can be expected in a mobile biometrics setting. Remarkably, during the on-stage demo, Face ID failed [1]. Immediate

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01267-0_16) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11219, pp. 263–281, 2018.
https://doi.org/10.1007/978-3-030-01267-0_16

speculation, especially from those with some familiarity with biometrics, centered around the possibility of a false negative, where an enrolled user failed to be recognized. After all, it was very dark on stage, with a harsh spotlight on the presenter, whose appearance was a bit more polished than usual—all variables that conceivably were not in the training set that the deep learning-based model behind Face ID was trained on. Apple, for its part, released a statement claiming that it was too many imposter authentication attempts before the demo that caused the problem [2]. Of course, that did little to satisfy the skeptics.

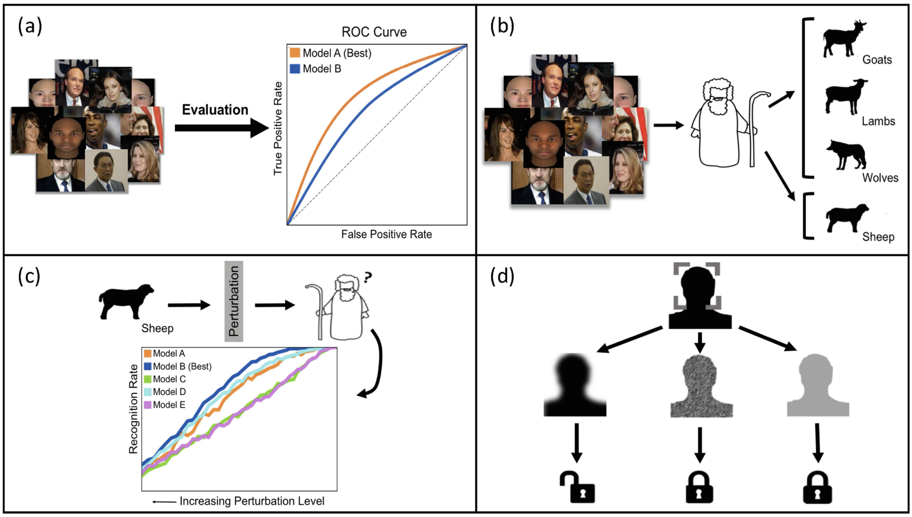

Fig. 1. Visual Pyschophysics [3–5] helps us explain algorithm behavior in a way that traditional dataset evaluation (a) cannot. Our proposed methodology introduces a theoretical mapping between elements of psychopysical testing and the biometric menagerie paradigm [6], where a shepherd function first isolates cooperative users ("sheep") from all others (b). From a perfect matching scenario, the images of the sheep are incrementally perturbed using a chosen image transformation, and item-response curves are plotted so that points of failure can be identified (c). The results can then be used to explain why matching works for some input images, but not others (d).

This controversy highlights a critical difficulty now facing the computer vision community: what is the true source of a problem when the object of study is a black box? While Apple may have access to the internals of its phones, ordinary users do not. But even with direct access to an algorithm, we can't always get what we want when it comes to an understanding of the conditions that lead to failure [7,8]. Given the fact that face recognition is one of the most common user-facing applications in computer vision, the ability to diagnose problems and validate claims about algorithm design and performance is desirable from the perspective of both the researcher and administrator charged with operating such systems. This is exactly why we want AI for face recognition to be *explainable*. In this paper, we look at a new methodology for doing this with any face recognition algorithm that takes an image as input. But first, let us consider the way we currently use evaluation procedures to try to understand their output.

The development cycle of face recognition algorithms relies on large-scale datasets. Progress is measured in a dataset context via summary statistics (*e.g.*, false positive rate, true positive rate, identification rate) computed over an evaluation set or n folds partitioned [9] from the evaluation set and expressed as a ROC or CMC curve (Fig. 1, Panel a). Such datasets have become even more important with the rise of machine learning, where both large training and evaluation sets are needed. For face verification (1:1 matching), there are a number of datasets that brought performance up to usable levels in controlled settings with cooperative subjects [10–14]. More recently, web-scale data [15–21] has been used to investigate more difficult recognition settings including face identification (1:N matching) and challenging impostor settings. There is a continuing push for larger datasets, which does not always address the problems observed in the algorithms trained over them. While aggregate performance has continued to improve, understanding individual causes of failure remains difficult, as it is not always clear why a particular face fails to be recognized, or why an impostor is recognized by an algorithm when considering a summary statistic.

Importantly, other fields studying vision have addressed this via the use of visual psychophysics: the controlled manipulation of stimuli and careful study of the responses they evoke in a model system [3–5]. In particular, the field of psychology has developed specific concepts and procedures related to visual psychophysics for the study of the human face and how it is perceived [22–25]. Instead of inferring performance from summary statistics expressed as curves like ROC or CMC, visual psychophysics allows us to view performance over a comprehensive range of conditions, permitting an experimenter to pinpoint the exact condition that results in failure. The gold standard for face recognition experimentation with people is the Cambridge Face Memory Test [23], which uses progressively degraded variations of faces to impede recognition. It has led to landmark studies on prosopagnosia (the inability to recognize a face) [26], super recognizers (people with an uncanny ability to recognize faces) [27], and face recognition ability and heritability [28]. Similarly, visual psychophysics has been used to study the role of holistic features in recognition by swapping parts to break the recognition ability [22]. More recent work has moved into the realm of photo-realistic 3D face synthesis, where changes in face perception can be studied by varying aspects of facial anatomy [24] and the age of the face used as a stimulus [25]. Given the breadth of its applicability, psychophysics also turns out to be an extremely powerful regime for explaining the behavior of algorithms.

We already see visual psychophysics becoming an alternate way of studying algorithm behavior in other areas of computer vision such as object recognition [29], face detection [30], and reinforcement learning [31]. However, no work has been undertaken yet in the area of face recognition. In this paper, we propose to address this by building a bridge from vision science to biometrics. Working from a recently established framework for conducting psychophysics experiments on computer vision algorithms [29] and infusing it with the proper methods from visual psychophysics for the study of face recognition in people, we fill in the missing pieces for automatic face recognition. Specifically, this involves a theoretical mapping between elements of psychopysical testing and the biometric menagerie paradigm [6], where cooperative users ("sheep") are isolated (Fig. 1, Panel b), and incremental perturbations degrade their performance (Fig. 1, Panel c). Results gathered from psychophysics experiments making use of highly controlled procedurally generated stimuli can then inform the way we should use a face recognition algorithm by explaining its failure modes (Fig. 1, Panel d).

2 Related Work

Explainable AI. An increasing emphasis on artificial neural networks in AI has resulted in a corresponding uptick in interest in explaining how trained models work. With respect to representations, Zeiler and Fergus [32] suggested that a multi-layer deconvolutional network can be used to project feature activations of a target convolutional network (CNN) back to pixel-space, thus allowing a researcher to reverse engineer the stimuli that excite the feature-maps at any layer in the CNN. Subsequent work by Mahendran and Vedaldi [33] generalized the understanding of representations via the analysis of the representation itself coupled with a natural image prior. With respect to decision making, Ribeiro et al. [8] have introduced a framework for approximating any classifier with an explicitly interpretable model. In a different, but related tactic, Fong and Vedaldi [34] use image perturbations to localize image regions relevant to classification. Image perturbations will form an important part of our methodology, described below in Sect. 3. A number of alternative regimes have also been proposed, including a sampling-based strategy that can be applied to face recognition algorithms [35], sampling coupled with reinforcement learning [7], and a comprehensive probabilistic programming framework [36]. What we propose in this paper is not meant to be a replacement for any existing method for explaining an AI model, and can work in concert with any of the above methods.

Psychophysics for Computer Vision. The application of psychophysics to computer vision has largely been an outgrowth of interdisciplinary work between brain scientists and computer scientists looking to build explanatory models that are consistent with observed behavior in animals and people. A recent example of this is the work of Rajalingham et al. [37], which compares the recognition behavior of monkeys, people and CNNs, noting that CNNs do not account for the image-level behavioral patterns of primates. Other have carried out studies using just humans as a reference point, with similar conclusions [38–41]. With respect to approaches designed specifically to perform psychophysics on computer vision algorithms, a flexible framework is PsyPhy, introduced by RichardWebster et al. [29]. PysPhy facilitates a psychophysical analysis for object recognition through the use of item-response theory. We build from that work to support a related item-response analysis for face recognition. Outside of research to explain the mechanisms of AI algorithms, other work in computer vision has sought to infuse psychophysical measurements into machine learning models [30,42]. Data in several of these studies has relied on the popular crowd-sourced psychophysics website TestMyBrain.org [43]. In this work, we make use of a similar human-testing platform for comparison experiments.

Methods from Psychology Applied to Biometrics. While there is growing interest in what psychology can teach computer vision at large, the biometrics community was early to adopt some of its methods. Sinha et al. [44] outlined 19 findings from human vision that have important consequences for automatic face recognition. Several of these findings have served as direct inspiration for the adoption of CNNs for face recognition. A significant outgrowth of NIST-run face recognition evaluations has been a series of human vs. computer performance tests [45–49]. Even though these studies have not made use of psychophysics, they still shed new light on face recognition capabilities. In some cases such as changes in illumination [45,46], good quality images [47], and matching frontal faces in still images [48], algorithms have been shown to be superior. However, one should keep in mind that these are controlled (or mostly controlled) verification settings, where images were intentionally acquired to reflect operational matching scenarios. In other cases, especially with more naturalistic data and video matching scenarios [48,49], humans are shown to be superior. Studies such as these have established human perception as a measureable baseline for evaluating face recognition algorithms. We also look at human vs. algorithm performance as a baseline in this paper.

Biometrics and Perturbed Inputs. Many studies have sought to simulate real-world conditions that reduce matching performance. This has often taken the form of perturbations applied to the pixels on a face image—the primary form of transformation we will consider for our psychophysics experiments. Karahan et al. [50] and Grm et al. [51] have studied the impact of incrementally perturbing face images for transformations like Gaussian blur, noise, occlusion, contrast and color balance. In order to compensate for Gaussian blur, Ding and Tao [52] perturb sequences of face images for the purpose of learning blur-insensitive features within a CNN model. These experimental studies share an underlying motivation with this work, but are qualitatively and quantitatively different from the item-response-based approach we describe.

3 Psychophysics for Face Recognition Algorithms

In the *M-alternative forced-choice match-to-sample* (M-AFC) psychophysics procedure in psychology [5], a *sample* stimulus (*e.g.*, visual, auditory, or tactile) is used to elicit a perceptual response from a subject. The subject is then given a refractory period to allow their response to return to neutral. Once their response returns to neutral, the subject is presented with an *alternate* stimulus and given, if needed, another refractory period. This process is then repeated for a total of M unique alternate stimuli. Finally, the subject is *forced* to choose one of the alternate stimuli that best *matched* the sample stimulus. This is where the procedure name M-alternative forced-choice match-to-sample comes from. By carefully linking sample or alternate stimuli to a single condition at a specific stimulus level, a scientist running the experiment can measure mean or median accuracy achieved at each of the observed stimulus levels across all subjects. Together, these stimulus levels and their aggregated accuracy yield an interpretable item-response curve [3] (see Fig. 1, Panel c for an example).

RichardWebster et al. [29] introduced a technique using the M-AFC method to produce item-response curves for general object classification models that involves procedurally rendering objects. The process consists of two steps: (1) the identification of a preferred view and (2) the generation of an item-response curve. A preferred view is an extension of a canonical view [53], the theory that humans naturally prefer similar inter-class object orientations when asked for the best orientation which maximizes discriminability. The preferred view serves as the initial orientation of the procedurally rendered objects, allowing transformations such as rotation or scaling to guarantee a degradation of model performance. When item-response curves are generated, a modified M-AFC procedure is invoked that maps the alternate choices to the output of a classifier. However, instead of explicitly presenting alternate choices, the alternate choices are implicitely the learned classes of the classifier. Thus accuracy is computed by how frequently the correct class was chosen.

Although psychophysics for face recognition uses the same foundational M-AFC match-to-sample concepts, in practice it is very different than the psychophysics procedure for general object recognition. To begin with, an individual trial of the M-AFC procedure described above for human subjects is identical to the face identification procedure of biometrics. A face is acquired, and the system is queried to determine the identity of the face by matching the acquired image to enrolled faces within the system. Thus, a single M-AFC match-to-sample trial is equivalent to 1:N identification in biometrics. However, one difference between an algorithm performing 1:N matching and a human performing the same task is the need to set a threshold for the decision of "match" or "non-match" in the case of the algorithm (to reject match instances with insufficiently high scores).

Like any good scientific method, a method from psychophysics attempts to isolate a single variable to observe the effect it has on the rest of the system. In psychophysics experiments for face recognition, we call the isolated variable the perturbation level, which represents the degree of transformation applied with a perturbation function directly to an identity or to the image containing an identity. Thus, the first step in performing psychophysics for face recognition systems is to remove identities from an initial dataset that consistently cause false matches or false non-matches—errors that are already inherent within the matching process and would be a confound to studying the effect of the transformation. Doddington et al. [54] formally grouped users interacting with a biometric system into four classes whimsically named after farm animals, which together are called the *biometric menagerie* [6,55]. The biometric menagerie consists of *goats* (identities that are difficult to match), *lambs* (identities that are easily impersonated), *wolves* (identities that impersonate easily), and finally *sheep* (identities that match well to themselves but poorly to others). Since we

Algorithm 1. $H(\Upsilon, I)$: a "herding" function to isolate Doddington et al.'s [54] sheep from the goats, lambs, and wolves

```
Input: $\Upsilon$, a "shepherd" function for a face recognition algorithm
Input: $I$, a set of input identities from a dataset
1: $S \leftarrow \Upsilon(I, I)$                                  ▷ similarity matrix
2: $S \leftarrow \frac{(S+S^{\intercal})}{2}$                              ▷ enforce symmetry
3: $t_h \leftarrow$ optimize loss function $\lambda$ with TPE             ▷ Hyperopt [56–58]
4: $I_h \leftarrow \lambda(S, t_h)$                 ▷ the "sheep" identities produced by $\lambda$
Output: $t_h$, the optimal threshold to produce $I_h$
Output: $I_h$, the "sheep" identities isolated by the optimal threshold $t_h$
```

want to remove all identities that lead to errors, we must remove the wolves, goats, and lambs. We call this the "herding" process.

The herding function, H (Algorithm 1), takes a set of input identities from an initial dataset, I, and a "shepherd" function, Υ, as input, and determines which identities Υ considers sheep. The Υ function is a wrapper function to a face recognition algorithm, f, and accepts two sets of identities: I_p the probe set and I_g the gallery set. It returns a standard similarity matrix where I_p is row-wise and I_g is column-wise. An example shepherd function can be seen in Algorithm 2. During the herding step, the input set I is split into I_p and I_g, which are used as input to Υ. The herding function itself is quite simple: it obtains a similarity matrix from the shepherd function, forces matrix symmetry, and then optimizes the loss function, λ (Algorithm 3), for 250 iterations with Hyperopt's implementation of the Tree-structured Parzen Estimator (TPE) hyperparameter optimizer [56–58]. More complicated is the loss function λ that the herding function uses.

λ takes as input a similarity matrix, S, and a threshold, t. The first step, thresholding the matrix, is standard in biometrics applications. However, the next step is not. The thresholded matrix is then XORed with an identity matrix, $\mathcal{I}$, to isolate all of the false match and false non-match pairs of identities ($\mathcal{I}$ represents the correct true matches). This new matrix can be considered an adjacency matrix, G, where all of the edges represent the false matches and false non-matches and each vertex is an identity.

The next step is to selectively remove vertices/identities until no edges remain while also removing as small a number of identities as possible. A strategy inspired by graph cuts allows us to sort the vertices by degree, remove the vertex with the highest degree from G, and repeat until no edges in G remain (see Supp. Algorithm 1 for the exact description[1]). At the end, G will be a completely disconnected graph, where no remaining identity will cause a false match or false non-match with any other remaining identity. By definition, all of the remaining identities are sheep. The returned loss value is the number of identities removed, where the function favors a lower false match rate, *i.e.*, higher thresholds are

[1] Supp. mat. available at http://www.bjrichardwebster.com/papers/menagerie/supp.

Algorithm 2. $\Upsilon_f(I_p, I_g)$: a "shepherd" function that produces a similarity matrix for the face recognition function f

```
Input: $f$, a face recognition function that produces a feature representation
Input: $I_p$, a set of probe identities
Input: $I_g$, a set of gallery identities
$R_p \leftarrow i \in I_p : f(i)$                              ▷ feature representation for each identity
$R_g \leftarrow i \in I_g : f(i)$
$S \leftarrow r_p \in R_p, r_g \in R_g : \mathrm{dist}(r_p, r_g)$              ▷ matrix of distances
$S \leftarrow \mathrm{normalize}(S)$                ▷ normalize distances to standard similarity matrix
Output: $S$, the similarity matrix
```

Algorithm 3. $\lambda(S, t)$: a loss function that favors more *sheep*, and favors a lower false match rate (FMR) over false non-match rate (FNMR)

```
Input: $S$, similarity matrix
Input: $t$, a threshold
$M \leftarrow S \geq t$
$M \leftarrow M \oplus \mathcal{I}$                                        ▷ isolate FM and FNM pairs
$G = (V, E)$ from $M$                                                  ▷ adjacency list
$\nu \leftarrow |V|$
while $|E| > 0$ do                                       ▷ remove goats, lambs, and wolves
    $v_r \leftarrow \mathrm{argmax}_{v \in V} \deg(v)$
    remove $v_r$ from $V$                   ▷ remove the vertex and connected edges from $G$
end while
$l \leftarrow \nu - |V|$                                ▷ number of goats, lambs, and wolves removed
$l \leftarrow l + (1 - 0.99999 * t)$                                 ▷ favor lower FMR over FNMR
Output: $l$, the loss value
```

favored. After λ is optimized, the optimal threshold t_h and sheep identities I_h are returned.

The sheep identities I_h and the threshold t_h serve as two of the inputs to the item-response point generator function Φ (Algorithm 4). Φ generates a point on an item-response curve that represents the rank one match rate for a specific perturbation function, T, and its respective perturbation level. The perturbation function takes an image and a perturbation level as input, applies some transformation to the image, and returns the transformed image. In the context of the biometric menagerie, this function is analogous to "perturbing" a sheep (dying the wool, shearing the wool, etc.) and asking its shepherd if it can properly identify the sheep. Thus Φ also takes Υ as a parameter. Φ uses T to perturb each input identity in I_h to create the set of perturbed probe identities for 1:N identification. The remaining steps of Φ are standard to face recognition systems operating in the identification mode: obtain similarity matrix from probe to gallery pairs, threshold the matrix, and calculate the match rate. The return value of the Φ function is an x, y coordinate pair $\{s, \alpha\}$ for one item-response point, where s represents the perturbation level and α is the match rate.

Algorithm 4. $\Phi_T(\Upsilon, I_h, t_h, \delta)$: an item-response point generation function for any image transformation function $T(i, \delta)$

```
Input: $\Upsilon$, a “shepherd” function for a facial recognition model
Input: $I_h$, the “sheep” identities for the found threshold $t_h$
Input: $t_h$, the optimal threshold to produce $I_h$
Input: $\delta$, the stimulus level
$I'_h \leftarrow i \in I_h : T(i, \delta)$                ▷ perturb identities to create probes
$S \leftarrow \Upsilon(I'_h, I_h)$                        ▷ similarity matrix
$M \leftarrow S \geq t_h$
$\alpha \leftarrow \frac{|M \wedge \mathcal{I}|}{|I_h|}$   ▷ obtain match rate using identity matrix $\mathcal{I}$
Output: $\{s, \alpha\}$, an $x, y$ coordinate pair (stimulus level, match rate)
```

Algorithm 5. $\mathcal{C}_T(\Upsilon, I_h, t_h, n, b_l, bu)$: an item-response curve generation function for any type of “shepherd” function

```
Input: $\Upsilon$, a “shepherd” function for a facial recognition model
Input: $t_h$, the optimal threshold to produce $I_h$
Input: $I_h$, the “sheep” identities for the found threshold $t_h$
Input: $n$, the number of stimulus levels
Input: $b_l$ and $b_u$, the lower and upper bound values of the stimulus levels
Let $\Delta$ be $n$ log-spaced stimulus levels from $b_l$ to $b_u$
$k \leftarrow \bigcup_{\delta \in \Delta} \{\Phi_T(\Upsilon, I_h, t_h, \delta)\}$
Output: $k$, the item-response curve
```

A shepherd's behavior for a set of sheep identities can be represented with an item-response curve (a collection of points obtained from Φ), which is an interpretable representation of the shepherd's behavior in response to perturbation. For biometric identification, the x-axis is a series of values that represent a perturbation level from the original sheep identities and the y-axis is the match rate. To produce the item-response curves, the function $\mathcal{C}$ (Algorithm 5) is called once for each transformation type. $\mathcal{C}$ repeatedly calls a point generated with Φ (Algorithm 4) to create one point for each stimulus level from the least amount of perturbation, b_l, to the most, b_u (b_l are the non-transformed sheep identities). The parameter n is the number of stimulus levels to be used to produce the points on the match-response curve and are typically log-spaced to give finer precision near the non-transformed sheep identities. The final parameter w is the number of identities examined at each stimulus level where $w \in [1, |I_h|]$.

4 Experiments

Experiments were designed with four distinct objectives in mind: (1) to survey the performance of deep CNNs and other alternative models from the literature; (2) to look more closely at a surprising finding in order to explain the observed model behavior; (3) to study networks with stochastic outputs, which are prevalent in Bayesian analysis; and (4) to compare human vs. algorithm performance.

For all experiments, we made use of the following face recognition algorithms: VGG-Face [59], FaceNet [60], OpenFace [61], a simple three-layer CNN trained via high-throughput search of random weights [62] (labeled “slmsimple” below), and OpenBR 1.1.0 [63], which makes use of handcrafted features. For each of the networks, the final feature layer was used with normalized cosine similarity as the similarity metric[2]. All used models were used as-is from their corresponding authors, with no additional fine-tuning. A complete set of plots for all experiments can be found in the supplemental material.

Data Generation. The following transformations were applied to 2D images from the LFW dataset [64]: Gaussian blur, linear occlusion, salt & pepper noise, Gaussian noise, brown noise, pink noise, brightness, contrast, and sharpness. Note that we intentionally chose LFW because state-of-the-art algorithms have reached ceiling performance on it. The psychophysics testing regime makes it far more difficult for the algorithms, depending on the chosen transformation. Each face recognition algorithm was asked to “herd” 1000 initial images before item-response curve generation. All algorithms except OpenBR recognized all the initial images as sheep (see Supp. Sect. 2 for a breakdown). For each transformation, we generated 200 different log-spaced stimulus levels, using each algorithm’s choice of sheep, to create a corresponding item-response curve. In all, this resulted in $\sim$5.5 *million* unique images and $\sim$13.7 *billion* image comparisons.

Inspired by earlier work in psychology [24,25,65] making use of the FaceGen software package [66], we used it to apply transformations related to emotion and expression. A complete list can be found in the supplemental material. Each face algorithm selected sheep from 220 initial images (all face textures provided by FaceGen, mapped to its “average” 3D “zero” model) for item-response curve generation. All chose 206 sheep, with a nearly identical selection by each (see Supp. Sect. 3 for a complete list). 50 stimulus levels were rendered for each image, resulting in $\sim 400,000$ unique 3D images and $\sim$17.5 *billion* image comparisons.

Identification with 2D Images. Given recent results on datasets, one would expect that the deep CNNs (FaceNet, OpenFace, and VGG-Face) would be the best performers on an M-AFC task, following by the shallower network (slmsimple), and then the approach that makes use of handcrafted features (OpenBR). Surprisingly, this is not what we observed for any of the experiments (Figs. 2 and 4; Supp. Figs. 1–2). Overall, VGG-Face is the best performing network, as it is able to withstand the perturbations to a greater degree than the rest of the algorithms. At some points (*e.g.*, left-hand side of Fig. 2) the perturbations have absolutely no effect on VGG-Face, while severely degrading the performance of other algorithms, signifying strong learned invariance.

Remarkably, the non-deep learning approach OpenBR is not the worst performing algorithm. It turned out to outperform several of the deep networks in most experiments. This is the kind of finding that would not be apparent from a CMC or ROC curve calculated from a dataset, where OpenBR is easily outperformed by many algorithms across many datasets [63,67]. Why does

[2] Source code is available at www.bjrichardwebster.com/papers/menagerie/code.

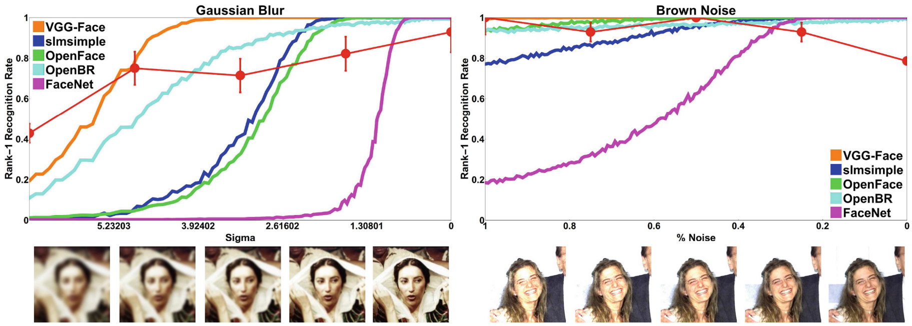

Fig. 2. A selection of item-response curves for the M-AFC task using data from the LFW dataset [64]. Each experiment used five different face recognition algorithms [59–63]. A perfect curve would be a flat line at the top of the plot. The images at the bottom of each curve show how the perturbations increase from right to left, starting with no perturbation (*i.e.*, the original image) for all conditions. The red dots indicate mean human performance for a selected stimulus level; error bars are standard error. Curves are normalized so chance is 0 on the y-axis. All plots are best viewed in color.

this occur? These results indicate that it's not always possible to rely on large amounts of training data to learn strongly invariant features—a task that can be different from learning representations that perform well on a chosen dataset. The design of the algorithm is also consequential: OpenBR's choice of LBP [68] and SIFT [69] leads to better performance than FaceNet and OpenFace, which each learned features over hundreds of thousands of faces images.

Identification with 3D Images. Computer graphics allows us to generate images for which all parameters are known—something not achievable with 2D data. One such parameter, expression, has been widely studied [70–72], but not in the highly incremental manner we propose here. Where exactly do algorithms break for specific expression changes? We can find this out by controlling the face with graphics (Figs. 3 and 4; Supp. Figs. 3–4). For instance, for the bodily function of blinking (Fig. 3) VGG-Face and slmsimple are the best, while this very small change to the visual appearance of the face causes a significant degradation of matching performance in the three other algorithms. OpenFace and FaceNet once again have trouble learning invariance from their training data. This trend holds over several expressions and emotions (Supp. Figs. 3–4).

OpenFace vs. FaceNet. It is often difficult to assess the claims made by the developers of machine-learning-based algorithms. During the course of our experimentation, we discovered an interesting discrepancy between two networks, FaceNet [60] and OpenFace [61], which both reported to be implementations of Google's FaceNet algorithm [20]. While it is good for end-users that deep learning has, in a sense, become "plug-and-play," there is also some concern surrounding this. It is not always clear if a re-implementation of an algorithm matches the original specification. Psychophysics can help us find this out. Across all experiments, FaceNet demonstrates very weak invariance properties compared to OpenFace (Figs. 3–4; Supp. Figs. 3–4), and fails well before the other algorithms

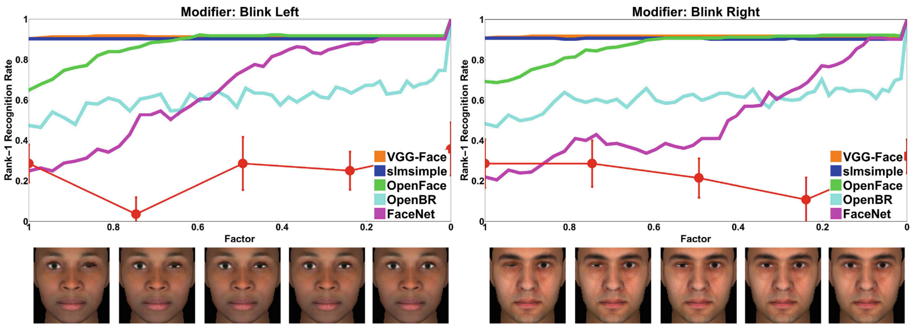

Fig. 3. A selection of item-response curves for the M-AFC task using rendered 3D face models as stimuli [66]. Curves are normalized so chance is 0. Here we see that three of the algorithms are drastically affected by the simple bodily function of blinking, while two others are not impacted at all. As in Fig. 2, VGG-Face is once again the best performing algorithm, but remarkably, we see that the three-layer CNN trained via a random search for weights (labeled "slmsimple") works just as well.

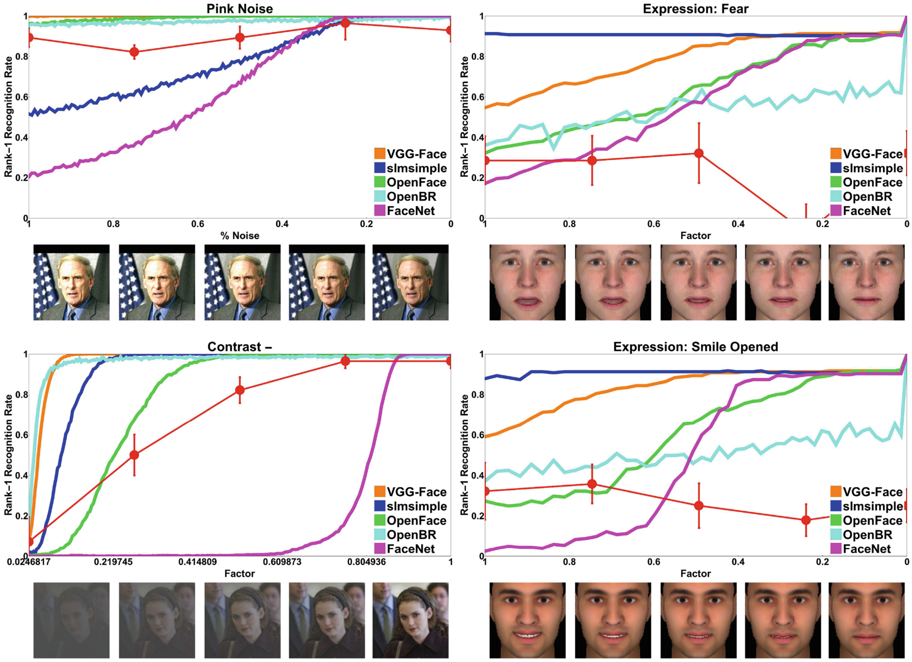

Fig. 4. Two of the algorithms we evaluated, FaceNet [60] and OpenFace [61], both reported to be an implementation of Google's FaceNet [20] algorithm. Curiously, we found major disagreement between them in almost all of our experiments. Note the gaps between their respective curves in the above plot. This performance gap was not evident when analyzing their reported accuracy performance on LFW.

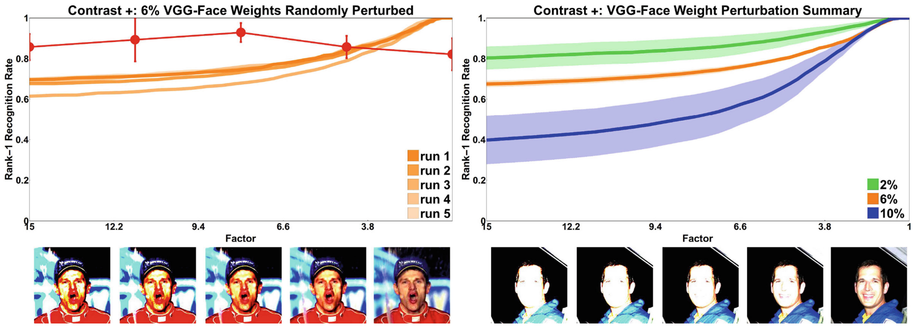

Fig. 5. Weight perturbations for stochastic model output can be combined with stimulus perturbations for a stronger reliability assessment. (Left) Five independent model runs where 6% of the weights have been perturbed, with input stimuli reflecting increasing contrast. (Right) Curves represent the average of five runs for three different levels of weight perturbation from 2% to 10%. Shaded regions are standard error.

in most cases. From these results, we can conclude that use of this particular implementation of Google's FaceNet should be avoided. But why is it so different from OpenFace, and what would be causing it to fail, in spite of it reporting superior accuracy on LFW (0.992 for FaceNet vs. 0.9292 for OpenFace)?

One can find three key differences in the code and data—after being prompted to look there by the psychophysics experiments. (1) OpenFace uses 500k training images by combining CASIA-WebFace [17] and FaceScrub [73]; FaceNet uses a subset of MS-Celeb-1M [74] where difficult images that contain partial occlusion, silhouettes, etc. *have been removed* as a function of facial landmark detection. This is likely the weakest link, as the network does not have an opportunity to learn invariance to these conditions. (2) OpenFace uses the exact architecture described by Schroff et al. [20], while FaceNet opts for Inception ResNet v1 [75]. (3) FaceNet uses a Multi-Task CNN [76] for facial landmark detection and alignment, while OpenFace uses dlib [77]—which FaceNet intentionally avoids due to its lower yield of faces for the training set. FaceNet may have hit upon the right combination of network elements for LFW, but it does not generalize like the original work, which OpenFace is more faithful to.

Weight Perturbation Coupled with Stimulus Perturbation. The procedure of applying perturbations directly to the weights of a neural network has an interpretation of Bayesian inference over the weights, and leads to stochastic output [78,79]. This is potentially important for face recognition because it gives us another measure of model reliability. To look at the effect of CNN weight perturbations coupled with stimulus perturbations, we use VGG-Face as a case study. A percentage of its weights are replaced with a random value from the normal distribution, $\mathcal{N}(0,1)$, targeting all layers. From Fig. 5, we can see that both perturbation types have an impact. Under a regime that perturbs just 6% of the weights (left-hand side of Fig. 5), we can gain a sense that VGG-Face is stable across models with respect to its performance when processing increasing levels of contrast. However, too much weight perturbation increases the variance, leading to undesirable behavior on the perturbed input. On the right-hand side of Fig. 5, each curve represents the average of five runs when perturbing between 2% and 10% of the weights. Perturbing 10% of the weights breaks the invariant features of VGG-Face and induces more variance between models. Similar effects for other transformations can be seen in Supp. Figs. 5–6.

Human Comparisons. As discussed in Sect. 2, there is a rich literature within biometrics comparing human and algorithm performance. However, thus far, such studies have not made use of any procedures from visual psychophysics. Here we fill this gap. To obtain human data points for Figs. 2–5 (the red dots in the plots), we conducted a study with 14 participants. The task the participants performed largely followed the standard M-AFC protocol described above: a participant is briefly shown an image, it is hidden from sight, and then they are shown three images and directed to choose the image that is most similar to the first one. Each participant performed the task three times for each perturbation level. Each set of images within a task was chosen carefully to keep human performance from being perfect. For both 2D and 3D images, the images were divided by gender such that participants could not match solely by it [80]. For 3D images, the data was also divided by ethnicity such that it could not be the sole criterion to match by [81]. To interrupt iconic memory [82], after each sample image is shown, a scrambled inverse frequency function was applied to the image to produce colored noise, and shown for 500ms prior to the alternate choices. 2D images were shown for 50ms and 3D images for 200ms. Humans struggled to identify faces in the 3D context where different identities are closer in visual appearance, but excelled in the 2D context where there was greater separation between identities. The plots for Gaussian blur (Fig. 2) and Decreasing Contrast (Fig. 4) hint at behavioral consistency between AI and humans in these cases.

5 Conclusion

Given the model capacity of today's deep neural network-based algorithms, there is an enormous burden to explain what is learned and how that translates into algorithm behavior. Psychophysics allows us to do this in a straightforward manner when methods from psychology are adapted to conform to the typical procedures of biometric matching, as we have shown. Companies launching new products incorporating face recognition can potentially prevent (or at least mitigate) embarrassing incidents like Apple's botched demo of FaceID by matching the operational setting of an algorithm to a useable input space. And even if a company provides an explanation for a product's failure, anyone can directly interrogate it via a psychophysics experiment to find out if those claims are true. To facilitate this, all source code and data associated with this paper will be released upon publication. With the recent uptick in psychophysics work for computer vision [29,30,37–40,42], we expect to see new face recognition algorithms start to use these data to improve their performance.

Acknowledgments. Funding was provided under IARPA contract #D16PC00002, NSF DGE #1313583, and NSF SBIR Award #IIP-1738479. Hardware support was generously provided by the NVIDIA Corporation, and made available by the National Science Foundation (NSF) through grant #CNS-1629914.

References

1. Apple Inc: IPHONE X FACE ID FAIL (2017). https://www.youtube.com/watch?v=m7xmCCTVS7Q. Accessed 1 Mar 2018
2. Hern, A.: Apple: Face ID didn't fail at iPhone X launch, our staff did. The Guardian, 14 Sep 2017
3. Embretson, S.E., Reise, S.P.: Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Inc., Mahwah (2000)
4. Lu, Z.L., Dosher, B.: Visual Psychophysics: From Laboratory to Theory. MIT Press (2013)
5. Kingdom, F., Prins, N.: Psychophysics: A Practical Introduction. Academic Press (2016)
6. Yager, N., Dunstone, T.: The biometric menagerie. IEEE Trans. Pattern Anal. Mach. Intell. **32**(2), 220–230 (2010)
7. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 3–19. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_1
8. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: explaining the predictions of any classifier. In: ACM KDD (2016)
9. Haralick, R.M.: Performance characterization in computer vision. In: Hogg, D., Boyle, R. (eds.) BMVC 1992, pp. 1–8. Springer, London (1992). https://doi.org/10.1007/978-1-4471-3201-1_1
10. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
11. Phillips, P.J., et al.: Overview of the face recognition grand challenge. In: IEEE CVPR (2005)
12. Phillips, P.J., et al.: Overview of the multiple biometrics grand challenge. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 705–714. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01793-3_72
13. Beveridge, J.R., et al.: The challenge of face recognition from digital point-and-shoot cameras. In: IEEE BTAS (2013)
14. Phillips, P.J., et al.: An introduction to the good, the bad, & the ugly face recognition challenge problem. In: IEEE FG (2011)
15. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: IEEE CVPR (2016)
16. Klare, B.F., et al.: Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In: IEEE CVPR (2015)
17. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch (2014). arXiv preprint arXiv:1411.7923
18. Ortiz, E.G., Becker, B.C.: Face recognition for web-scale datasets. Comput. Vis. Image Underst. **118**, 153–170 (2014)

19. Bhattarai, B., Sharma, G., Jurie, F., Pérez, P.: Some faces are more equal than others: hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8926, pp. 160–172. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16181-5_12
20. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: IEEE CVPR (2015)
21. Wang, D., Otto, C., Jain, A.K.: Face search at scale. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1122–1136 (2017)
22. Tanaka, J.W., Farah, M.J.: Parts and wholes in face recognition. Q. J. Exp. Psychol. **46**(2), 225–245 (1993)
23. Duchaine, B., Nakayama, K.: The Cambridge face memory test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. Neuropsychologia **44**(4), 576–585 (2006)
24. Oosterhof, N.N., Todorov, A.: The functional basis of face evaluation. Proc. Natl. Acad. Sci. **105**(32), 11087–11092 (2008)
25. Germine, L.T., Duchaine, B., Nakayama, K.: Where cognitive development and aging meet: face learning ability peaks after age 30. Cognition **118**(2), 201–210 (2011)
26. Duchaine, B., Germine, L., Nakayama, K.: Family resemblance: ten family members with prosopagnosia and within-class object agnosia. Cogn. Neuropsychol. **24**(4), 419–430 (2007)
27. Russell, R., Duchaine, B., Nakayama, K.: Super-recognizers: people with extraordinary face recognition ability. Psychon. Bull. Rev. **16**(2), 252–257 (2009)
28. Wilmer, J.B., et al.: Human face recognition ability is specific and highly heritable. Proc. Natl. Acad. Sci. **107**(11), 5238–5241 (2010)
29. RichardWebster, B., Anthony, S., Scheirer, W.: Psyphy: a psychophysics driven evaluation framework for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 1–1 (2018). Preprint
30. Scheirer, W.J., Anthony, S.E., Nakayama, K., Cox, D.D.: Perceptual annotation: Measuring human vision to improve computer vision. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1679–1686 (2014)
31. Leibo, J.Z., et al.: Psychlab: a psychology laboratory for deep reinforcement learning agents (2018). arXiv preprint arXiv:1801.08116
32. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
33. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: IEEE CVPR (2015)
34. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: IEEE ICCV (2017)
35. Turner, R.: A model explanation system. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP) (2016)
36. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
37. Rajalingham, R., Issa, E.B., Bashivan, P., Kar, K., Schmidt, K., DiCarlo, J.J.: Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks, 240614 (2018). bioRxiv

38. Gerhard, H.E., Wichmann, F.A., Bethge, M.: How sensitive is the human visual system to the local statistics of natural images? PLoS Comput. Biol. **9**(1), e1002873 (2013)
39. Eberhardt, S., Cader, J., Serre, T.: How deep is the feature analysis underlying rapid visual categorization? In: NIPS (2016)
40. Geirhos, R., Janssen, D.H.J., Schütt, H.H., Rauber, J., Bethge, M., Wichmann, F.A.: Comparing deep neural networks against humans: object recognition when the signal gets weaker (2017). arXiv preprint arXiv:1706.06969
41. Heath, M., Sarkar, S., Sanocki, T., Bowyer, K.: Comparison of edge detectors: a methodology and initial study. In: 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Proceedings CVPR 1996, pp. 143–148. IEEE (1996)
42. McCurie, M., Beletti, F., Parzianello, L., Westendorp, A., Anthony, S.E., Scheirer, W.J.: Predicting first impressions with deep learning. In: IEEE FG (2017)
43. Germine, L., Nakayama, K., Duchaine, B.C., Chabris, C.F., Chatterjee, G., Wilmer, J.B.: Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. Psychon. Bull. Rev. **19**(5), 847–857 (2012)
44. Sinha, P., Balas, B., Ostrovsky, Y., Russell, R.: Face recognition by humans: nineteen results all computer vision researchers should know about. Proc. IEEE **94**(11), 1948–1962 (2006)
45. O'Toole, A.J., Phillips, P.J., Jiang, F., Ayyad, J., Penard, N., Abdi, H.: Face recognition algorithms surpass humans matching faces over changes in illumination. IEEE Trans. Pattern Anal. Mach. Intell. **29**(9), 1642–1646 (2007)
46. O'Toole, A.J., Phillips, P.J., Narvekar, A.: Humans versus algorithms: comparisons from the face recognition vendor test 2006. In: IEEE FG (2008)
47. O'Toole, A.J., An, X., Dunlop, J., Natu, V., Phillips, P.J.: Comparing face recognition algorithms to humans on challenging tasks. ACM Trans. Appl. Percept. (TAP) **9**(4), 16 (2012)
48. Phillips, P.J., O'Toole, A.J.: Comparison of human and computer performance across face recognition experiments. Image Vis. Comput. **32**(1), 74–85 (2014)
49. Phillips, P.J., Hill, M.Q., Swindle, J.A., O'Toole, A.J.: Human and algorithm performance on the PaSC face recognition challenge. In: IEEE BTAS (2015)
50. Karahan, S., Yildirum, M.K., Kirtac, K., Rende, F.S., Butun, G., Ekenel, H.K.: How image degradations affect deep CNN-based face recognition? In: International Conference of the Biometrics Special Interest Group (BIOSIG) (2016)
51. Grm, K., Struc, V., Artiges, A., Caron, M., Ekenel, H.K.: Strengths and weaknesses of deep learning models for face recognition against image degradations. IET Biom. **7**(1), 81–89 (2018)
52. Ding, C., Tao, D.: Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2017, to appear)
53. Blanz, V., Tarr, M.J., Bülthoff, H.H.: What object attributes determine canonical views? Perception **28**(5), 575–599 (1999)
54. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Technical report, National Institute of Standards and Technology (1998)
55. Teli, M.N., Beveridge, J.R., Phillips, P.J., Givens, G.H., Bolme, D.S., Draper, B.A.: Biometric zoos: theory and experimental evidence. In: IEEE/IAPR IJCB (2011)
56. Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: NIPS (2011)

57. Bergstra, J., Yamins, D., Cox, D.D.: Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms. In: 12th Python in Science Conference (2013)
58. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., Cox, D.D.: Hyperopt: a python library for model selection and hyperparameter optimization. Comput. Sci. Discov. **8**(1), 014008 (2015)
59. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC, vol. 1, p. 6 (2015)
60. Sandberg, D.: Face recognition using tensorflow (2017). https://github.com/davidsandberg/facenet. Accessed 1 Mar 2018
61. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: a general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science (2016)
62. Cox, D.D., Pinto, N.: Beyond simple features: a large-scale feature search approach to unconstrained face recognition. In: IEEE FG (2011)
63. Klontz, J.C., Klare, B.F., Klum, S., Jain, A.K., Burge, M.J.: Open source biometric recognition. In: IEEE BTAS (2013)
64. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report, 07–49, University of Massachusetts, Amherst (2007)
65. Yildirim, I., Kulkarni, T.D., Freiwald, W.A., Tenenbaum, J.B.: Efficient and robust analysis-by-synthesis in vision: a computational framework, behavioral tests, and modeling neuronal representations. In: Annual Conference of the Cognitive Science Society (2015)
66. Singular Inversions: Facegen (2017). https://facegen.com/. Accessed 1 Mar 2018
67. Amos, B.: Openface 0.2.0: higher accuracy and halved execution time (2016). http://bamos.github.io/2016/01/19/openface-0.2.0/. Accessed 1 Mar 2018
68. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
69. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE ICCV (1999)
70. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: IEEE FG (2002)
71. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image Vis. Comput. **28**(5), 807–813 (2010)
72. Dutta, A., Veldhuis, R., Spreeuwers, L.: A Bayesian model for predicting face recognition performance using image quality. In: IEEE/IAPR IJCB (2014)
73. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: IEEE ICIP (2014)
74. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 87–102. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_6
75. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
76. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
77. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)

78. Graves, A.: Practical variational inference for neural networks. In: NIPS (2011)
79. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT Press Cambridge (2016)
80. O'Toole, A.J., Deffenbacher, K.A., Valentin, D., McKee, K., Huff, D., Abdi, H.: The perception of face gender: the role of stimulus structure in recognition and classification. Mem. Cogn. **26**(1), 146–160 (1998)
81. Webster, M.A., MacLeod, D.I.: Visual adaptation and face perception. Philos. Trans. R. Soc. Lond. B Biol. Sci. **366**(1571), 1702–1725 (2011)
82. Dick, A.O.: Iconic memory and its relation to perceptual processing and other memory mechanisms. Percept. Psychophys. **16**(3), 575–596 (1974)